Synopsis to be submitted for the Award of Degree of Doctor of Philosophy in Computer Science

Exploring Learning Algorithms for Classification of Dysphonia using Speech Signals

By
Sharal Coelho
Research Scholar
(MU/EXB/Ph.D/CR.39/Com.Sc./2022-23/E.25)

Under the Guidance of
Dr. H. L. Shashirekha
Professor

Post Graduate Department of Studies and Research in Computer Science
Mangalore University, Mangalagangothri, Konaje
Karnataka, India
October 2025

Contents

1 Introduction **4**

1.1 Voice Disorders: Definitions, Causes, and AI-Based Detection 4

1.1.1 Preprocessing: 6

1.1.2 Acoustic Features 7

1.2 Publicly Available Datasets for Diagnostic Study 9

1.3 Limitations of Existing Methods 10

1.4 Motivation 10

1.5 Challenges in Dysphonia Classification 10

1.6 Objectives 11

2 Proposed Work 1: KanVD – Creation of Speech Corpus for Identifying Voice Disorder/Dysphonia **12**

2.1 Related Work 13

2.2 Recording Equipment 14

2.3 Protocol 14

2.4 Statistics 15

2.5 Data Organization 15

2.6 Baselines 16

2.6.1 Data Pre-processing 16

2.6.2 Feature Extraction 16

2.6.3 Model Building 17

2.7 Experiments and Results 17

2.7.1 Results 18

2.7.2 Error Analysis 19

2.8 Summary 21

3 Proposed Work 2: Classification of Dysphonia with Acoustic Feature Sets **21**

3.1 Related Work 21

3.2 Methodology 21

3.2.1 Data Pre-processing and Feature Extraction 22

3.2.2 Classification Algorithms 22

3.3 Experiments 22

3.4 Results and Discussion 23

3.4.1 Feature Set Comparison 23

3.5 Comparative Evaluation on the SVD Dataset 24

4 Proposed Work 3: Classification of Dysphonia using Balancing Techniques 25
4.1 Related Work . . . 25
4.2 Methodology . . . 26
4.2.1 Pre-processing . . . 27
4.2.2 Feature Extraction . . . 27
4.2.3 Model Building . . . 27
4.3 Experiments . . . 28
4.3.1 Dataset Description . . . 29
4.4 Results and Discussion . . . 30

5 Proposed Work 4: Classification of Voice Disorder using Spectrograms of Voice Signals 31
5.1 Related Work . . . 32
5.2 Methodology . . . 32
5.2.1 Feature Representation . . . 33
5.2.2 Feature Concatenation . . . 34
5.2.3 Classification Models . . . 34
5.3 Experiments . . . 35
5.3.1 Dataset Description . . . 35
5.4 Results and Discussion . . . 36

6 Proposed Work 5: A Semi-supervised Co-training Strategy for Automated Voice Disorder Detection 37
6.1 Related Work . . . 37
6.2 Methodology . . . 38
6.2.1 Data Preparation and Feature Extraction . . . 38
6.2.2 Model Architecture . . . 39
6.2.3 Semi-supervised Learning using Pseudo-Labeling . . . 39
6.3 Experiments . . . 40
6.3.1 Dataset Description . . . 40
6.4 Results and Discussion . . . 41
6.5 Comparative Study . . . 41

7 Proposed Work 6: Automatic Assessment of Pathological Voice Quality using Learning Approaches 42
7.1 Dataset Description . . . 43
7.2 Related Work . . . 43
7.3 Methodology . . . 44
7.3.1 Pre-processing and Feature Extraction . . . 44
7.3.2 Model Construction . . . 45

7.4 Experimental Results and Discussion 46

8 Conclusion and Future Works **49**

9 Publications **49**

9.1 Publications in International Conferences 49

9.2 Papers Presented in National Conferences 50

9.3 Papers Under Preparation 50

10 Organization of the Thesis **50**

1 Introduction

1.1 Voice Disorders: Definitions, Causes, and AI-Based Detection

Communication is an active process of exchanging information and ideas. All human activities require better communication and the most effective way of communication is through speech [1]. Speech signals provide information about the spoken words, speaker identity [2], language spoken [3], physical and mental health, age, and sex [4]. While voice is the sound produced by the larynx, speech is more of a mechanical act of movement of the articulators - tongue, lips, teeth, and hard palate, to produce a series of vowel and consonant sounds. The process of speech production is depicted in Figure 1. When air is pushed to the vocal folds with sufficient pressure in human beings, the vocal folds vibrate resulting in the production of voice that includes talking, laughing, crying, yelling, etc.

A voice disorder/pathology is any condition that affects the ability to produce a normal voice, often impairing communication and reducing quality of life [5]. This can involve changes in pitch, loudness, tone, or vocal quality. People with voice disorders may sound hoarse, breathy, strained, or may lose their voice entirely. Voice disorders are common among individuals who use their voice heavily in their profession, such as teachers, singers, actors, and public speakers.

Figure 1: Human Speech Production

There are several common types of voice disorders. Laryngitis is one of the most frequent, caused by inflammation of the vocal cords due to infection or overuse. Vocal nodules and polyps are growths on the vocal cords resulting from repeated strain, like yelling or prolonged

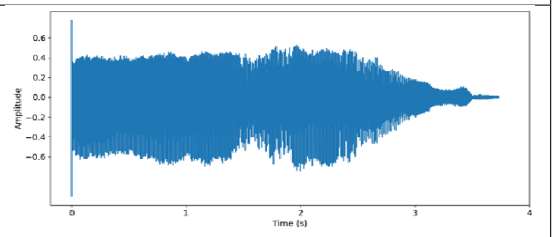
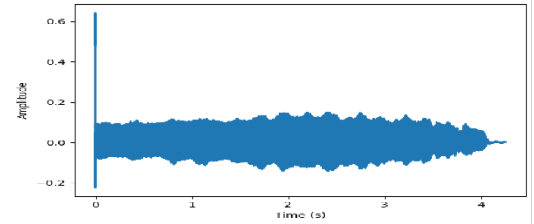
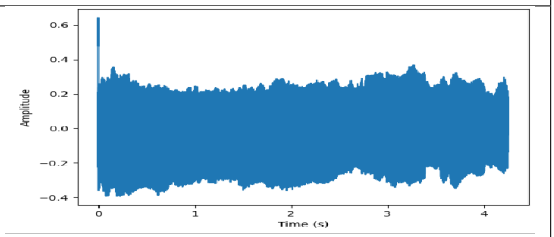
Condition	Causes	Waveform
Laryngitis	Overuse, irritation, or infection	
Vocal Cord Nodules	Singing, yelling, or talking loudly	
Muscle Tension Dysphonia	Increased muscle tension in muscles surrounding the voice box, often linked to high stress levels	

Table 1: Common Voice Disorders: Causes and Phonatory Waveform Patterns

speaking. Spasmodic dysphonia is a neurological condition where involuntary muscle spasms affect the vocal cords, making speech sound jerky or shaky [6]. Vocal cord paralysis happens when one or both vocal cords cannot move, often due to nerve damage, surgery, or injury [7]. Lastly, functional voice disorders occur when there's improper use of the voice without any physical damage to the vocal cords. Such disorders may arise from structural, neurological, or functional abnormalities in the vocal mechanism, including the larynx and vocal folds. Waveform representations for selected voice disorders along with causes are illustrated in the Table 1.

The consequence of voice disorders can be significant. Socially, they can make communication difficult and frustrating [8]. Emotionally, many people experience stress, humiliation, or depression when they are unable to speak normally or comfortably. Professionally, they can limit or even end careers in voice-dependent fields. Early identification and accurate diagnosis of voice disorders are critical for effective treatment and prevention of long-term complications. Traditional diagnostic methods, such as laryngoscopy and perceptual evaluation, though effective, are often subjective, time-consuming, and require specialized clinical expertise.

With the advancement of digital signal processing and Machine Learning (ML) tech-

niques, automated systems for Voice Disorder Detection / Dysphonia Detection have gained attention. These systems analyze acoustic signals to extract relevant features and classify them into normal or disordered voice categories. Such tools offer potential for non-invasive, cost-effective, and scalable screening, especially in remote or resource-limited environments. This study aims to explore the methodologies, challenges, and future directions in the development of automated systems for Voice Disorder Ientification, with a focus on integrating signal processing and Artificial Intelligence (AI) techniques. The simple framework for the ML approach for voice disorder classification process is illustrated in Figure 2 and the steps includes (i) Preprocessing: Preparing and resampling audio samples for consistency, (ii) Feature Extraction: Extracting relevant features from audio data for analysis, and (iii) Classification in Target Classes: Implementing the ML model for accurate classification.

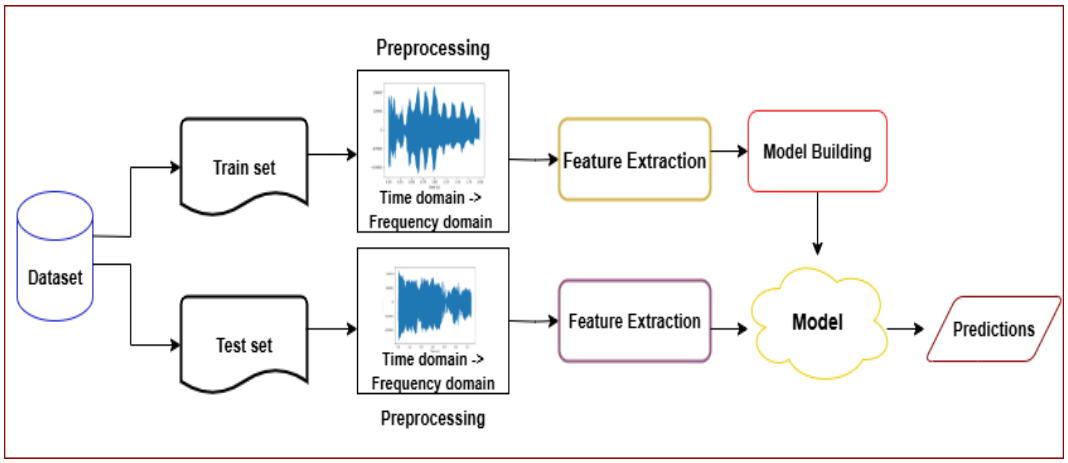


Figure 2: Framework of Machine Learning approach for Voice Disorder Classification

1.1.1 Preprocessing:

In the pre-processing stage, resampling plays a critical role in preparing audio signals for analysis by converting them from one sample rate to another, such as from 44 kHz to 22 kHz (downsampling) or from 16 kHz to 22 kHz (upsampling). This time-domain operation adjusts the number of samples per second, directly impacting the frequency content available for subsequent frequency-domain analysis. Downsampling, for instance, reduces the sample rate by discarding samples, which may result in the loss of high-frequency components, while upsampling increases the sample rate by interpolating new samples. By transforming the audio signal from the time domain to the frequency domain, resampling enables the analysis of its frequency components, ensuring compatibility with downstream processing requirements.

1.1.2 Acoustic Features

AI models analyze acoustic signals by extracting features from voice recordings [9]. Some of the key voice/acoustic features used in AI-based diagnosis include:

- **Mel-Frequency Cepstral Coefficients (MFCCs)** captures the changes in vocal tract configuration and reflects the perceptual characteristics of voice and has commonly been used in speaker identification, voice disorder identification, and speech recognition systems [10]. The 13 MFCC coefficients remain as standard features due to their ability to capture the spectral form of speech without requiring a lot of computer resource.
- **Jitter** measures the cycle-to-cycle variation in frequency, indicating instability in pitch [11].
- **Shimmer** reflects the variation in amplitude between consecutive vocal cycles, signaling irregular vocal fold vibration [11].
- **Harmonics-to-Noise Ratio (HNR)** indicates the overall periodicity of the voice signal by quantifying the ratio between the aperiodic (noise) and periodic (harmonic part) components. It is measured as an overall characteristic of the signal, and not as a function of frequency [12].
- **Pitch and Energy** help to identify voice range, prosody, and strength.
- **Mel spectrogram** is a spectrogram where the frequencies are converted to the mel scale. It is widely used in speech processing and analysis[13].
- **Chroma** features identify the pitch details of the speech signals affected by the voice disorders due to disruptions or irregularities in the voice generation process aiding in the identification of abnormalities.
- **Spectral contrast** is useful for detecting spectral changes in speech signal and is often used for classifying voice disorders [14].
- **Spectral centroid** indicate the spectral energy distribution in detecting disturbance in voice production [15].
- **Root-Mean-Square (RMS)** value provides a measure of the signal's energy, helping to detect irregularities or inconsistencies in the voice that may indicate the presence of a disorder.
- **Zero-Crossing Rate (ZCR)** detects irregularities in the speech signal by calculating the rate at which the audio signal changes sign from positive to negative or vice versa. Abnormal vocal patterns often result in different ZCR compared to normal speech [16].

- **Spectral roll-off** is commonly used to distinguish between harmonic content and noisy content or fricative sounds. In the context of identifying voice quality, spectral roll-off helps to detect abnormal frequency distributions in the voice signal that may indicate the changes in the actual voice [17].

These features can be used with ML classifiers like Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), etc. Few of them are explained below:

- **Decision Tree** is a hierarchical classification model which splits data into branches based on feature values and each path from root to leaf node represents a decision rule based on the features and leaf nodes represent predictions [18].
- **Bagging Classifer (BC)**[1] is an ensemble meta estimator that fits each base classifier (DT) on random subsets of the original dataset before combining by voting or averaging to produce the final prediction.
- **Gradient Boosting (GB)** is a robust boosting classifier that integrates several weak learners into strong learners and training new model to minimize the loss function.
- **eXtreme Gradient Boosting (XGB)** which belongs to the family of boosting algorithms is an ensemble learning technique to combine the predictions of multiple weak learners.
- **Random Forest** algorithm utilizes a large number of self-learning DTs. For each DT, RF first generates a random sample of the training dataset and the attributes are then randomly selected at each node to establish the optimum rule for separating the data and eventually selecting a class label [19].
- **Logistic Regression** is a ML classifier used to predict categorical dependent variables [20]. While working with multi-class classification problem LR is instantiated with multi_class = 'multinomial' indicating that it will perform multinomial LR using the softmax function and solver 'lbfgs' is a used for optimizing model.
- **ExtraTreesClassifier (ETC)** is an ensemble approach used to build a classification model. This model trains numerous DTs and aggregates the results from the group of DTs to output a prediction.

While working with Deep Learning (DL) models, the features extracted are taken as input, especially Convolutional Neural Network (CNN) [21] [22], when voice signals are represented as spectrograms. Spectrograms visualize frequency content over time and have proven

[1]https://www.geeksforgeeks.org/ml-bagging-classifier/

especially effective in DL-based voice disorder classification tasks [23][24]. The combination of signal-based and image-based representations has opened new frontiers in developing robust, real-time, and accurate voice disorder detection systems. These methods allow the system to learn complex patterns in voice signals that are not easily captured by traditional handcrafted features. DL models, especially, can automatically extract deep acoustic features and handle variations across speakers, making them ideal for voice pathology detection [10].

Despite the growing interest in automated voice disorder identification systems, several limitations still hinder their widespread clinical application. One major challenge is the variability in voice data due to differences in language, accent, age, sex, and recording conditions. Additionally, publicly available datasets for voice disorders are limited in size and diversity, which affects the generalizability of trained models. Most systems also struggle with real-time performance and the accurate classification of borderline or mixed disorders.

1.2 Publicly Available Datasets for Diagnostic Study

Using audio data, the learning approaches were trained for diverse applications such as detecting and monitoring voice pathologies, speech recognition[25], analyzing voice data, and diagnosing voice pathologies [26]. However, for all these experiments there is a need of large amount of data which is crucial in developing computer based tool to identify the voice disorder. Currently, there are few databases of voice samples existing in literature and often they are not easily available. One of the most used databases is the Massachusetts Eye and Ear Infirmary (MEEI). It contains voice recordings of subjects with a wide variety of voice disorders and also normative subjects voice samples. Even though it is considered as the basis of many studies of voice pathology assessment, it has some limitations, such as it is not freely available for the scientific community, and that different environments and sample frequencies are used to record the healthy and pathological voices [27].

The "VOice ICar fEDerico II (VOICED)" voice dataset has been introduced to help researchers who needs voice signals in their research activities. It VOICED dataset is publically available on PhysioNet website and the dataset consists of pathological voice samples including three different pathologies (hyperkinetic and hypokinetic dysphonia and reflux laryngitis) and healthy voices in the data set [28]. The summarised details are shown in in Table 2.

Saarbruecken Voice Database (SVD) is a German database maintained by the "Institute of Phonetics" of Saarland University, Germany. It contains voice recordings that address 71 different pathologies such as Laryngitis, Vocal polyp, and Reinke Edema. The signals are sampled at 50 kHz with a 16-bit resolution [29]. A recording session contains the recordings of the vowels (/a/,/i/, and /u/) produced at normal, high, low and rising-falling pitch and recording of a sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?")

Table 2: The summarized details of available voice disorder dataset

Dataset	Recorded Samples	P	H	Total
VOICED [28]	Recordings of vowel /a/	150	58	208
MEEI [27]	Recordings of vowel /a/	657	53	710
	Utterance of the first sentence of the rainbow passage	662	53	715
SVD [30]	Recordings of /a/, /i/, /u/ and rainbow passage	1354	687	2041

P - Pathology, H - Healthy

1.3 Limitations of Existing Methods

Despite technical advancements, several challenges persist in the domain of voice disorder classification. The lack of large, balanced, and diverse datasets, especially from Indian populations, is a major limitation. Most studies rely on Western datasets [31], which may not generalize well to the Indian population due to linguistic and phonetic differences. Furthermore, many existing models are trained only on sustained vowels or scripted speech, limiting their applicability in real-world scenarios [32]. Variability in recording environments and hardware further complicates the model's robustness.

1.4 Motivation

The motivation behind this study arises from the need to improve early detection and monitoring of voice disorders through efficient, reliable, and accessible technologies. Given the limitations of traditional methods and the potential of ML in healthcare, there is a significant opportunity to develop a system that can analyze vocal patterns and accurately identify disorders with minimal clinical supervision. By utilizing signal processing techniques and learning algorithms, we aim to develop a model that not only detects voice disorders but also provides interpretable results that may help in clinical decisions. This work aims to contribute to making vocal health assessment more inclusive.

In summary, the objective of this proposal is to develop a Kannada Voice Disorder (KanVD) Dataset corpus of sustained vowel samples to distinguish between (i) normal and pathological voice samples and (ii) to identify the quality of Dysphonia using learning algorithms. The outcome of this study is expected to help the Doctors/Clinicians/Speech therapists to detect the voice pathological disorder more effectively at a much lesser time.

1.5 Challenges in Dysphonia Classification

Several challenges are identified in the field of Dysphonia classification. They are as follows:

- **Data Collection:** Voice samples of Dysphonia may not be available at one source and hence, need to be collected from multiple sources like hospitals, speech and hearing institutions, etc. These samples, collected from different sources may not use the same strategy to record the voice, making the Dysphonia classification very challenging.

- **Unavailability of Data:** Many researchers exploring ways to detect Dysphonia depend on a small private dataset because of the non-availability of a larger dataset. Although several databases for voice pathology detection are available, databases for voice pathology specific to Kannada-speaking participants' voice samples are not available. Using voice pathological databases of other languages, such as English (i.e.,MEEI), Arabic(i.e., AVPD), and German (i.e., SVD) [27], to train the models to identify the voice pathologies in the local population, especially with Kannada-speaking people, does not give good results. The way Kannada-speaking people produce sound is different from how it is produced in other countries. Hence, it is necessary to have a large pathological voice database of Kannada-speaking population that contains voice samples of various types of voice disorders and also samples from healthy people.

- **Data Annotation:** Annotating the voice samples requires the services of experts/speech pathologists, which is expensive and laborious. This manual process involves careful auditory-perceptual evaluations using standardized scales like the GRBAS or CAPE-V [33], demanding focused listening and subjective decisions that vary between raters. The need for multiple annotators to achieve inter-rater reliability is slowing down research timelines significantly.

- **Ethical and Legal Challenges:** The major challenge in collecting data. Most people are unwilling to share recordings of their voice due to reasons such as shyness, privacy concerns, and security issues. Additionally, it is very important to obtain permission from a human ethics committee if collecting healthcare data, which again takes time to follow through with the necessary documentation [34].

These observations have motivated us to propose the classification of voice disorders using speech samples, with the aim of improving the performance of classification systems by experimenting with different features, feature extraction methods, and learning algorithms. In light of this, the following objectives are formulated for the proposed work.

1.6 Objectives

After an extensive review of the available literature and the previously done research work, we have come across certain gaps. Below given are some of the objectives that are in our focus :

- **Kannada Voice Disorder (KanVD) Dataset: Corpus Creation of Dysphonia Voice Samples** : The sustained vowel samples of patients who have Dysphonia/voice disorder are recorded by expert clinicians in a sound-treated room in speech and hearing centers and hospitals. In addition, voice samples from participants who have a healthy voice have to be collected after their clinical evaluation to make sure that they are not suffering from Dysphonia or any voice-related problems. The main aim is to develop a speech corpus, especially samples from people who speak Kannada.

- **Classification of Dysphonia using Speech Signals:** The aim is to distinguish between normal and Dysphonia voice samples using different learning algorithms such as SVM, kNN, DT, and CNN. The initial stage comprises of extracting significant features from the collected voice samples. Different feature extraction techniques will be used to extract jitter, shimmer, MFCC, etc. depending on their importance in accurately identifying the pathology.

- **Classification of Dysphonia Voice Quality using Voice Samples :** The aim of this objective is to identify the voice quality such as hoarse, breathy, harsh and normal from the voice samples using different learning algorithms. This can be modeled as multi-class classification problem.

2 Proposed Work 1: KanVD – Creation of Speech Corpus for Identifying Voice Disorder/Dysphonia

This study aims to introduce a newly developed voice dataset and evaluate its effectiveness in training models for automatic voice disorder detection. This study was conducted in adherence with all institutional and regulatory requirements. Ethical clearance was obtained from the Yenepoya Ethics Committee-2 (YEC-2), and approval was secured from the Scientific Review Board (SRB) before initiating the study. Additionally, a Memorandum of Understanding (MoU) was executed between Mangalore University, Mangalore and Yenepoya Medical College, Mangalore, to facilitate data sharing and research activities.

The collected dataset consists sustained vowel samples (/a/,/i/, and /u/) from Kannada speakers, including individuals with diagnosed voice disorders. The dataset includes recordings from varied age groups (18-70) and sex (Male and Female). Signal processing techniques are used to extract voice features such as MFCC. These features are fed into ML models for classification. The below subsections contains related work, the steps to designing and developing the KanVD dataset which includes recording equipment and protocol, an overview of the collection of samples, the statistics of the dataset, and the baselines. Figure 3 summarizes the steps undertaken during the data collection process.

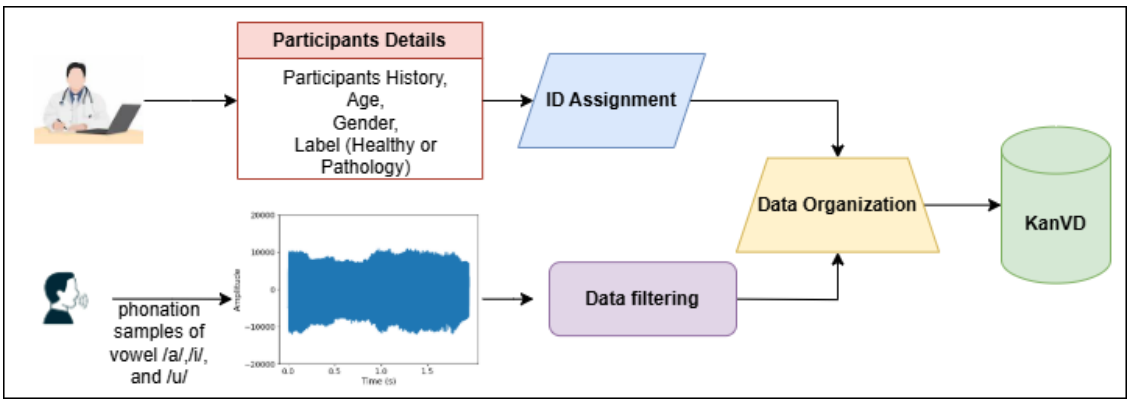


Figure 3: Procedures followed for data collections and data filtering

2.1 Related Work

The studies on voice disorder detection and classification are becoming increasingly important due to increased voice disorder issues in human population. Research works in the area of voice disorder detection and classification started in early 1980s [35]. Since then, research and experimentation have continuously evolved from traditional ML approaches to DL, and more recently, to transfer learning, paving the way for ongoing advancements in the field [36]. While some studies have progressed based on using different representations of speech samples, some have addressed the situation of limited data for classification tasks. Most existing studies rely on publicly available datasets and they often only contain data for specific disorders in particular languages [37].

Verma et al. [38] proposed a novel approach named VDDMFS (Voice Disorder Detection using MFCC, Fundamental Frequency, and Spectral Centroid) using VOICED dataset. The approach is developed by combining an Artificial Neural Network (ANN) trained on acoustic features with a Long Short-Term Memory (LSTM) model trained on MFCC features. The output probabilities from both models are then fused and passed to an XGBoost classifier to differentiate the presence of a voice disorder. This approach achieved superior performance, reporting an accuracy of 95.67%, sensitivity of 95.36%, specificity of 96.49%, and an F1-score of 96.9%. AL-Dhief et al.[39] has introduced an automatic voice pathology detection system utilizing the MFCC features and SVM classifier. The assessment of the introduced system was conducted based on voice signals of the vowel /a/ which been taken from SVD database. The study initially used 160 voice samples (80 healthy and 80 pathological) and progressively expanded the dataset to 280 samples, maintaining an equal distribution. The results have demonstrated that the uppermost attained accuracy by the SVM classifer was 84.37% using 160 voice utterances.

Although numerous studies have been conducted on voice disorder detection system using AI, there is a noticeable gap in research specifically focused on the Indian context– particularly in regions like Karnataka. Most existing studies rely on publicly available

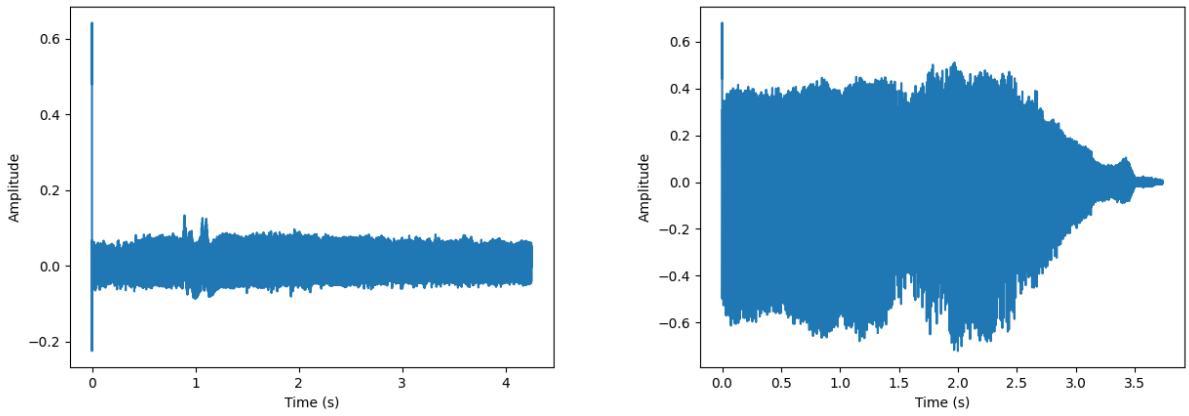
datasets, often collected in foreign languages, rather than utilizing data reflective of local linguistic and demographic diversity.

2.2 Recording Equipment

The phonation samples of the vowels /a/, /u/, and /i/ are collected from participants using Computerized Speech Lab model 4500 (CSL 4500) in a sound-treated room. While some participants provided their recordings via CSL, others used Praat. To maintain uniformity, the recording settings are kept identical across both tools, with a sampling rate of 44 kHz and a resolution of 16-bit. Further, samples that are not recorded properly or have noticeable issues are excluded from the dataset. Healthy and pathological voice signal samples from the KanVD dataset are shown in Figure 4.

2.3 Protocol

As part of the data collection process, Clinicians/SLPs asked participants a series of preliminary questions about their age, gender, profession, smoking habits, relevant medical history, and other speech-related complaints. Further, the participants were briefed about the purpose of collecting their voice samples followed by asking them to produce the vowel sounds /a/, /u/, and /i/. Each vowel was sustained comfortably at a normal pitch for a minimum duration of 3 seconds. The recordings were saved in .wav format, and additional trials were conducted if any sample was not properly captured. To ensure the reliability of the normal/healthy class, the voice samples from healthy participants were collected after a clinical evaluation confirmed that they do not have any symptoms of voice disorders.



(a) (b)

Figure 4: Voice signals of (a) Healthy (b) Pathology voice samples

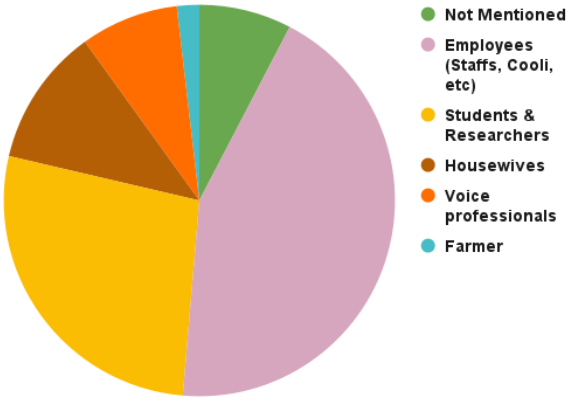


Figure 5: Distribution of the study population according to occupational status

2.4 Statistics

KanVD dataset includes 281 and 367 participants with normative and pathological voices, respectively. Each of them has a different occupational status (Student, Researcher, Housewife, Farmer, voice professional such as Singer, Teacher, etc). Pathological voice samples correspond to 32 different voice disorders, including Vocal fold nodules, Vocal cysts, Muscle Tension Dysphonia, Sulcus vocalis. Each participant is assigned a unique ID and from each participant, sustained voice samples of vowels /a/, /i/, and /u/, are collected resulting in a total of (281 + 367) X 3 = 1,944 samples. Table 3 shows the statistics of gender-wise distribution of participants and distribution of data according to occupational status is shown in Figure 5.

Table 3: Distribution of KanVD dataset

KanVD	**Male**	**Female**	**Total**
Healthy	126	155	281
Pathology	185	182	367
Total	311	337	648

2.5 Data Organization

The dataset includes the following details about each participant and is compiled as a .csv file:

- Participant ID: Numerical value serves as an identifier for the participant.
- Gender: Female or Male

- Age: In years
- Occupational status: Student, Farmer, Researcher, Housewife, Singer, Teacher, etc,
- Smoking habits: Yes or No
- Sample: Sustained recordings of the vowels /a/, /i/, and /u/.
- Class: Healthy (Annotated as 0) and Pathological (Annotated as 1)

This structured format allows for efficient data management, ensuring that participant-specific attributes are preserved.

2.6 Baselines

To evaluate the KanVD dataset, various experiments for classification of voice disorders are performed by using Learning approaches. The methodology to build models that automatically classify voice samples as healthy or pathological consists of: pre-processing, feature extraction, and model building.

2.6.1 Data Pre-processing

As a preprocessing step, minimal work is performed as the majority of preprocessing occurs during data collection and organization, where critical tasks such as noise removal are executed to ensure data quality. For instance, if a recording is considered impaired due to distortions or artifacts, it is re-recorded to meet quality standards. Additionally, this phase involves converting NSP files to WAV format to facilitate compatibility with analysis tools. The only preprocessing task isolated to this step is the conversion of the audio sample rate, which is adjusted to standardize the dataset for further processing and analysis.

2.6.2 Feature Extraction

One of the most prevalent ways to get acoustic features for automatic voice disorder detection systems is using the MFCC feature extraction technique [40]. MFCC features are extracted by applying the following steps:

- Pre-processing: For pre-emphasis, framing, and windowing of the voice signal.
- Fast Fourier Transform (FFT): Short-time Fourier analysis is performed to obtain the magnitude spectrum.
- Mel Filterbank: To convert the spectrum to Mel scale using 24 triangular filters.
- Log Power: For each filter bank output.
- Discrete Cosine Transform (DCT): To obtain 13 MFCC coefficients.

Table 4: Parameters of ML classifiers and their values

Classifier	Hyperparameter
AdaBoost	learning_rate=0.5, n_estimators=150
GB	n_estimators=100,learning_rate=0.1
RF	max_depth=10
ExtraTrees	n_estimators=250
Bagging	max_samples=0.7, n_estimators=50
XGB	learning_rate=0.1, max_depth=4, n_estimators=200, random_state=None
LR	max_iter=1000
kNN	n_neighbors=7, weights='distance'
SVM	C=1.0, kernal=rbf
DT	criterion='entropy', max_depth=10, min_samples_leaf=5

Table 5: Performances of ML models in Exp-1 using 3-fold cross-validation

Model	Precision	Recall	F1-Score	Accuracy
LR	64.89	64.80	64.61	65.07
DT	64.74	64.86	64.67	65.12
AdaBoost	66.70	66.21	66.15	66.82
GB	71.68	71.55	71.55	72.07
BC	73.55	73.56	73.54	73.97
kNN	74.31	74.67	74.30	74.49
RF	74.48	74.30	74.35	74.85
XGB	74.61	74.66	74.60	75.00
SVM	76.94	77.11	76.99	77.31
ExtraTrees	**77.87**	**77.49**	**77.63**	**78.14**

2.6.3 Model Building

The extracted MFCC features are used to train various ML classifiers (AdaBoost, GB, RF, ExtraTrees, BC, XGB, LR, kNN, SVM, and DT). Table 4 shows ML models, parameters and their values used in the experiments.

2.7 Experiments and Results

In order to establish the performances of baselines on KanVD dataset, the experiments are divided into two major categories:

1. Experiment 1 (Exp-1): all the sustained phonations /a/, /i/, and /u/, are considered for training and testing.

2. Experiment 2 (Exp-2): learning models are trained and tested individually with each

Table 6: Performances of ML models in Exp-1 using 5-fold cross-validation

Model	Precision	Recall	F1-Score	Accuracy
AdaBoost	66.81	66.11	66.17	67.13
DT	68.75	68.65	68.32	68.67
LR	68.41	67.90	67.90	68.77
GB	73.63	73.30	73.35	73.92
BC	75.25	74.60	74.70	75.36
RF	75.96	75.12	75.26	75.97
kNN	75.93	76.15	75.92	76.23
XGB	76.95	76.46	76.52	77.05
SVM	78.63	78.22	78.26	78.75
ExtraTrees	**80.11**	**78.96**	**79.20**	**79.89**

vowel - /a/, /i/, and /u/. This facilitates the identification of each vowel's discriminative capacity and its impact on classification performance.

Experiments are conducted using 3-fold Cross-Validation (3-fold CV) and 5-fold Cross-Validation (5-fold CV), to ensure robust evaluation and generalization of the models.

2.7.1 Results

To assess the performance of the ML models, precision, recall, F1-score, and accuracy are employed as the primary evaluation metrics. The performances of the ML models in Exp-1 using 3-fold and 5-fold CV are shown in Tables 5 and Table 6 respectively. Among all ML models, ExtraTrees model achieved the best accuracy of 78.14% and 79.89% for 3-fold CV and 5-fold CV, respectively. The performances of the classifiers in Exp-2 using 3-fold and 5-fold CV are shown in Table 7 and Table 8, respectively.

The results indicate that the ExtraTrees and SVM classifiers achieved the highest accuracies of 82.56% and 79.63%, respectively, using 3-fold CV with vowel /a/ samples. For 5-fold CV, the same models achieved accuracies of 81.32% and 80.55%, respectively. ExtraTrees classifier outperformed SVM classifier in both 3-fold and 5-fold CV, with the highest accuracy of 82.56% in the 5-fold CV. The cross-validation technique provides a more robust estimate of performance across multiple data subsets, making ExtraTrees more reliable and acceptable in this study. These models serve as robust baselines for KanVD dataset, establishing a strong foundation for further analysis and optimization.

Table 7: Performances of ML models in Exp-2 using 3-fold cross-validation

Model	Vowel	Precision	Recall	F1-Score	Accuracy
DT	a	72.05	72.00	71.97	72.53
	i	65.69	65.84	65.26	65.43
	u	68.41	68.52	68.38	68.83
AdaBoost	a	75.51	75.77	75.52	75.77
	i	69.96	69.75	69.58	70.06
	u	66.25	65.68	65.71	66.36
GB	a	77.27	77.49	77.34	77.62
	i	74.01	74.25	74.07	74.38
	u	73.05	73.18	73.08	73.46
kNN	a	77.75	77.88	77.68	77.93
	i	76.02	76.24	75.88	76.08
	u	73.76	74.08	73.60	73.77
RF	a	79.42	79.44	79.42	79.78
	i	76.53	76.56	76.5	76.85
	u	74.06	73.64	73.66	74.23
BC	a	79.28	79.48	79.34	79.63
	i	74.49	74.62	74.51	74.85
	u	73.35	73.09	73.01	73.61
XGB	a	78.93	78.99	78.95	79.32
	i	74.94	75.11	74.97	75.31
	u	73.87	73.90	73.77	74.23
LR	a	78.61	78.69	78.54	78.86
	i	71.15	71.14	70.83	71.14
	u	64.37	64.25	64.09	64.51
SVM	a	79.37	79.34	79.29	79.63
	i	80.14	80.18	80.09	80.40
	u	75.66	75.76	75.70	76.08
ExtraTrees	**a**	**82.43**	**81.89**	**82.06**	**82.56**
	i	**78.92**	**78.72**	**78.76**	**79.17**
	u	**76.19**	**75.95**	**76.03**	**76.54**

2.7.2 Error Analysis

Although the ExtraTrees classifier demonstrated a highest accuracy of 82.56%, it exhibits misclassifications due to certain characteristics. The dataset has intrinsic ambiguity in some recordings, i.e. healthy voice samples with mild variations acoustically resembled disordered voice leading to misclassifications. These recordings resulted in overlapping feature patterns in the MFCC space, which impacted the model's ability to recognize the classes correctly.

Table 8: Performances of ML models in Exp-2 using 5-fold cross-validation

Model	Vowel	Precision	Recall	F1-Score	Accuracy
DT	a	76.19	75.85	75.73	76.23
	i	66.13	66.17	65.80	66.20
	u	68.41	68.52	68.38	68.83
AdaBoost	a	77.6	77.24	77.26	77.77
	i	71.73	71.52	71.45	72.07
	u	66.25	65.68	65.71	66.36
GB	a	79.35	79.15	79.11	79.47
	i	75.06	74.97	74.88	75.30
	u	73.05	73.18	73.08	73.46
RF	a	79.80	79.25	79.27	79.78
	i	79.18	78.32	78.42	79.00
	u	74.06	73.64	73.66	74.23
BC	a	78.66	78.34	78.22	78.70
	i	75.35	75.18	74.91	75.30
	u	73.35	73.09	73.01	73.61
XGB	a	79.85	79.39	79.46	79.93
	i	77.46	76.75	76.77	77.31
	u	73.87	73.90	73.77	74.23
LR	a	77.51	76.83	76.70	77.31
	i	74.48	74.24	74.29	74.84
	u	64.37	64.25	64.09	64.51
kNN	a	80.41	80.39	80.24	80.54
	i	76.57	76.52	76.36	76.68
	u	73.76	74.08	73.60	73.77
SVM	**a**	80.64	80.10	80.10	80.55
	i	81.14	80.35	80.50	81.01
	u	75.66	75.76	75.70	76.08
ExtraTrees	**a**	**81.50**	**80.61**	**80.73**	**81.32**
	i	**81.67**	**80.50**	**80.71**	**81.32**
	u	**76.19**	**75.95**	**76.03**	**76.54**

Further, DT and GB models yielded lower accuracy compared to XGB and ExtraTrees, due to several factors such as model limitations and experimental settings. DT tends to underfit as it lacks the robustness to capture complex patterns in the data. Although GB is more capable, it is sensitive to noise and often requires careful hyperparameter tuning to perform well. In this study, a relatively lower number of estimators is used for both DT and GB, which further restricts their learning capacity. In contrast, XGBoost employs regularization and advanced optimization strategies that improve generalization, while ExtraTrees influences randomization and ensemble learning to handle variability more effectively.

2.8 Summary

This study presents the development of the KanVD - a voice signal dataset consisting of sustained vowel samples of /a/, /i/, and /u/. This dataset is used to develop and evaluate various ML models for the classification of voice disorders. Among all ML models, Extra-Trees demonstrated better performance in terms of accuracy, precision, recall, and F1-score. The findings show that KanVD dataset can be used to develop automated voice disorder classification systems to serve as effective decision-support tools for clinicians, potentially improving diagnostic accuracy and patient outcomes.

3 Proposed Work 2: Classification of Dysphonia with Acoustic Feature Sets

The proposed study investigates the impact of different voice features extracted from a KanVD dataset for the classification of voice disorders. The all experiments are conducted using various decision tree-based classifiers including DT, RF, XGBoost, and Gradient Boosting on different feature sets extracted from a KanVD dataset to evaluate their effectiveness in classifying voice disorders.

3.1 Related Work

Several studies have explored the use of various acoustic features and experimental setups for Dysphonia detection. Researchers have explored several datasets (SVD, MEEI [27], and Arabic Voice Pathology Database (AVPD) [41]), several voice features (spectral roll-off, spectral centroid [42], spectral contrast, chroma, MFCCs [43], and ZCR) and several classifiers (RF [44], kNN [45], SVM [46], and Gaussian Mixture Models (GMM) [47]) for identifying voice pathologies.

Traditional features such as MFCC, pitch, jitter, and shimmer have been widely used to characterize vocal attributes [35] [38]. Researchers have also experimented with feature selection techniques like PCA, LDA, and mutual information to enhance classifier performance [48]. On the experimental front, both classical ML algorithms (e.g., SVM, RF, GB) and DL architectures (e.g., CNNs, LSTMs) have been investigated. These diverse methods highlight the importance of feature engineering and model selection in achieving robust classification performance in Dysphonia detection tasks.

3.2 Methodology

This study aims to automatically identify voice disorders/Dysphonia using different set of acoustic features derived from sustained vowel phonations.

3.2.1 Data Pre-processing and Feature Extraction

Initially, all phonation samples /a/,/i/, and /u/ are resampled to a consistent sampling rate using the librosa library. Then three sets of features are extracted:

1. **Spectral Features** : Spectral centroid, Spectral bandwidth, Spectral contrast, Spectral flatness, and Spectral roll-off are the spectral features used for the initial experiments to evaluate their standalone performance in classifying the audio samples.

2. **Praat-Based Acoustic Features**:

   - Jitter Measures: Jita, Jitt, RAP, PPQ, and DDP
   - Shimmer Measures: Shim, ShdB, APQ3, APQ5, and DDA
   - Other Parameters: HNR (Harmonics-to-Noise Ratio) and pitch values.

3. **Extended Feature Representation using Mel Spectrogram** To capture more detailed time-frequency characteristics of the audio signal, Mel spectrogram features are later extracted. The Mel spectrogram represents the short-time power spectrum of the sound, mapped onto the Mel scale. For each audio sample, the Mel spectrogram is computed and averaged across the time axis, yielding a 128-dimensional feature vector representing energy distribution across Mel-frequency bands. This provided a more robust feature representation compared to just using spectral roll-off, ZCR and other features.

3.2.2 Classification Algorithms

In this work, classification is performed using decision tree-based algorithms due to their interpretability and effectiveness in handling structured data. The models employed include the standard DT classifier, along with its ensemble variants: RF, XGBoost, and GB. These models are evaluated using different extracted feature sets to determine their suitability for binary classification of voice disorders.

3.3 Experiments

The experiment is conducted on KanVD dataset which consists of Vowel /a/,/i/ and /u/. Each phonation samples annotated with metadata including age, sex, and Label (Healthy or Pathology). The implementation is carried out in Python, utilizing several key libraries: librosa is used for extracting Spectral features, parselmouth served as a Python interface to Praat for extracting acoustic voice parameters, and scikit-learn is employed for implementing and evaluating various ML classifiers.

The experiments are conducted in three set of voice features:

Table 9: Performance Comparison of Tree-Based Classifiers for Voice Disorder Detection with Feature sets

Feature Set	**Model**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Accuracy (%)**
Praat features only	DT	78.47	74.55	76.46	74.04
	RF	79.3	81.82	80.54	77.63
	GB	81.25	82.73	81.98	79.43
	XGB	**82.73**	**82.73**	**82.73**	**80.46**
Spectral features only	DT	78.28	70.45	74.16	72.24
	RF	**83.41**	**77.73**	**80.47**	**78.66**
	GB	80.38	76.36	78.32	76.09
	XGB	81.95	76.36	79.06	77.12
Spectral features + Mel spectrogram	DT	83.56	85.45	84.49	82.26
	RF	92.31	92.73	92.52	91.52
	GB	90.17	95.91	92.95	91.77
	XGB	**91.7**	**95.45**	**93.54**	**92.54**

- Phase 1: Using Praat features
- Phase 2: Using Spectral features
- Phase 3: Using combination of Spectral features and Mel spectrogram

The feature vectors are normalized prior to training using standard scaling. A stratified 80:20 train-test split is used to preserve class balance.

3.4 Results and Discussion

Multiple ML algorithms are evaluated for the task of classifying Healthy versus pathological voices. To assess the performance of the models, a comprehensive set of evaluation metrics is used and the results are shown in Table 9. These included accuracy, which measures the overall correctness of the model; precision, indicating the proportion of correctly predicted positive instances among all predicted positives; recall, which reflects the model's ability to identify all relevant instances; F1-score, the harmonic mean of precision and recall, which evaluates the classifier's ability to distinguish healthy and pathology classes. These metrics provided a well-rounded evaluation of the system's effectiveness in detecting Dysphonia.

3.4.1 Feature Set Comparison

Three experimental conditions are compared in this study– (i) using Spectral features only, (ii) using Praat features only, (iii) using combination of spectral and Mel spectrogram features. This comparison is designed to evaluate the individual effectiveness of spectral features

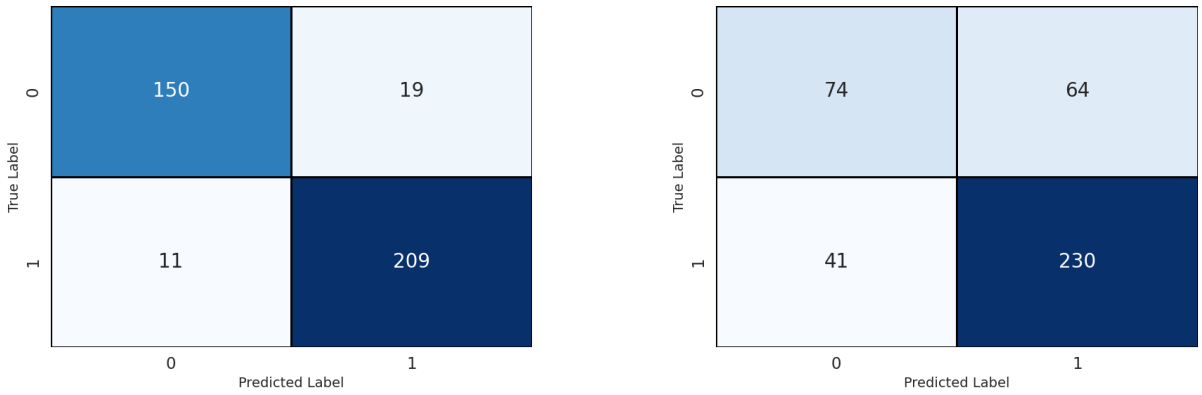
versus traditional perturbation and voice quality measures (extracted using Praat) in the classification of Dysphonia. By isolating the two feature sets, the experiment aimed to determine whether traditional acoustic measures such as jitter, shimmer, and HNR could complement or rival the performance of spectral features in distinguishing pathological voices. This analysis helped assess the relative contribution of each feature type and their potential synergy in enhancing classification accuracy.

The initial results using only spectral features are modest, suggesting the need for richer features. Incorporating the Mel spectrogram significantly improved classification performance. This methodology highlights the progressive improvement in performance as richer spectral representations and Mel spectrograms are introduced.

The performance of each classifier under different feature settings is summarized below. The combined feature set consistently outperformed individual sets across classifiers. The experimental results indicate that XGBoost outperformed other tree-based models when trained on the combined feature set, yielding an accuracy of 92.54%, along with 91.7% precision and 95.45 % recall values.

3.5 Comparative Evaluation on the SVD Dataset

The proposed model yielded superior performance on the third feature set, prompting us to apply the same methodology to the SVD dataset for comparative analysis. This extension also produced enhanced results, as detailed in Table 10. Furthermore, the corresponding confusion matrces is illustrated in Figure 6 below.



(a) (b)

Figure 6: Confusion matrices showing the performance of the model on (a) KanVD dataset and (b) SVD

The enhanced performance observed on our dataset (e.g., a 18% improvement in accuracy, as detailed in Table 10) highlights its efficacy as a benchmark for the proposed methodology. The corresponding confusion matrix further illustrates the model's improved handling of

Table 10: Comparative analysis of proposed study

Feature Set	Model	Proposed study	Proposed Methodology on SVD Data
Spectral features + Mel spectrogram	DT	82.26	62.84
	RF	91.52	71.88
	GB	91.77	72.13
	XGB	**92.54**	**74.33**

class imbalances, with reduced false positives in underrepresented classes. This advantage may be explained by the incorporation of diverse real-world scenarios, which are not present in the SVD dataset.

4 Proposed Work 3: Classification of Dysphonia using Balancing Techniques

The goal of this study is to distinguish between healthy and pathological voice samples by extracting the spectral features (MFCCs, ZCR, spectral roll-off, spectral centroid, spectral contrast, mel spectrogram, and chroma) from sustained samples of the vowel /a/, to train the ML models (AdaBoost, Bagging classifier, GB, Multi-layer Perceptron (MLP), kNN, LR, DT, RF, and XGBoost). The experiments are carried out on the subset of SVD [49] created by selecting the voice samples of vowel /a/ at normal, low, high, and low-high-low pitches of healthy and few voice disorders (Psychogene Dysphonie, Hyperfunktionelle Dysphonie, Laryngitis, Reinke –dem, Dysarthrophonie). The distinctive characteristics of the voice samples of these disorders help to distinguish between healthy and pathological voice disorders. The subset of the SVD dataset created consists of more pathological samples than healthy samples making the dataset imbalanced and imbalanced data may affect the performance of the classifier. Hence, the data imbalance problem is addressed by resampling techniques (Synthetic Minority Oversampling Technique (SMOTE), Borderline-SMOTE, Adaptive Synthetic Sampling (ADASYN), and SMOTE-Edited Nearest Neighbor (SMOTE-ENN)).

4.1 Related Work

From audio signals and EGG signals of the SVD dataset, Omeroglu et al. [46] extracted hand-crafted features (MFCC, Linear Predictive Coefficients (LPC), F0 (pitch), spectral slope) and using CNN they extracted deep features to identify the voice disorders. To obtain the subset of features from the collection of deep and other features they applied the InfoGainAttributeEval[2] feature selection method. They trained the SVM model using the

[2]https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html

subset of features and obtained an accuracy of 90.10%. Fan et al. [50] has addressed the class imbalance problem of speech samples in the MEEI database using SMOTE, Borderline-SMOTE, and ADASYN. In their other work Fan et al. [51] proposed a Fuzzy Clustering (FC)-SMOTE algorithm to detect and classify different types of pathological voices in a class-imbalanced database that outperforms SMOTE, Borderline-SMOTE, and ADASYN. For binary classification, the RF model obtained recall and specificity of 1.00 and 1.00 respectively. Due to the dataset's imbalance VOICED datasset, Arslan et al. [52] developed a model that offers automated speech pathology detection systems to improve diagnostic performance. The synthetic minority oversampling technique (SMOTE) is used to create the class-balanced data on VOICED dataset samples.

According to the related work, different learning approaches including ML and DL have employed classification metrics like accuracy, which can give misleading results when applied to imbalance datasets. Further, the performance of many ML models are not explored in the identification of voice disorders, implying the need to explore models for voice disorders to improve the performance.

4.2 Methodology

The proposed methodology for the identification of voice disorders using sustained samples of vowel /a/ includes: i) Pre-processing, ii) Feature Extraction, and iii) Model Building. The framework of the ML approach is shown in Figure 7 and a description of the steps are given below:

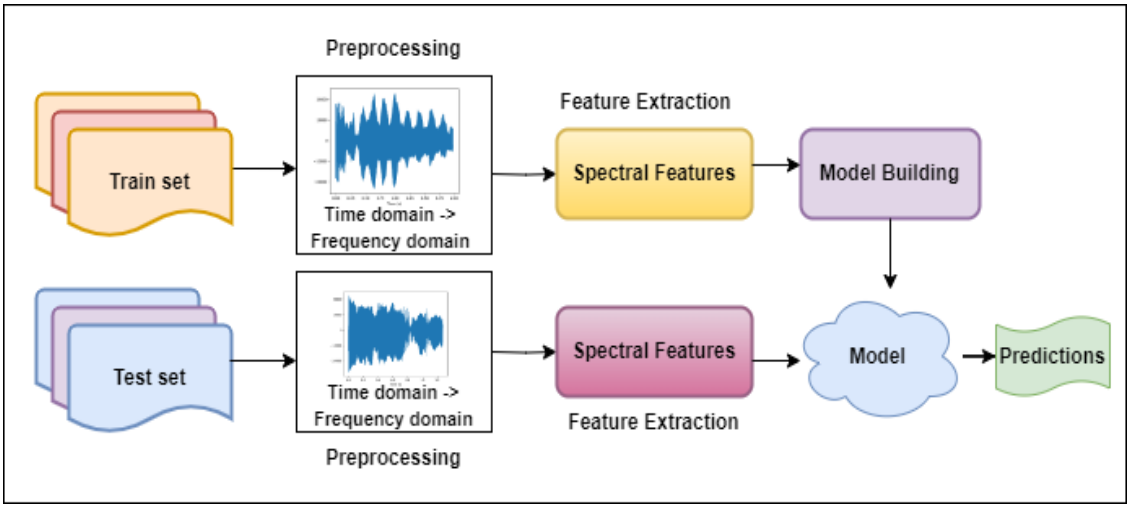


Figure 7: Framework of Machine Learning approach

4.2.1 Pre-processing

The sustained vowel samples are converted from time domain to frequency domain as speech signals in the frequency domain provide frequency distribution of the speech signals which is important for examining the frequency components to detect voice disorders. If the audio samples have higher sample rate, they are resampled to a lower sample rate as lowering the sample rate decreases the amount of data that is required to be processed and analyzed.

4.2.2 Feature Extraction

The pathological voice samples are characterized by more irregular spectral properties as compared to healthy voice samples and these irregularities can be captured by the spectral features. Spectral roll-off, spectral centroid, spectral contrast, mel spectrogram, chroma, MFCCs, and ZCR features are the spectral features extracted from the given voice signals. MFCC is a spectral feature that captures irregularities or variations in the spectral properties of the voice signals from voice signals even if the voice samples consist of noise. Spectral centroid captures the changes in an abnormal voice by providing the single value of the center of mass of the speech signal's spectrum [53]. Spectral roll-off focuses on the distribution of energy in the frequency domain. Spectral contrast is useful for detecting spectral changes in speech signal and is often used for classifying voice disorders [14]. ZCR represents the rate at which a signal changes its sign from positive to negative or vice versa over time.

4.2.3 Model Building

AdaBoost, Bagging classifier, GBoosting, DT, MLP, kNN, LR, RF, and XGBoost models are explored for the identification of voice disorders. AdaBoost is an ensemble ML algorithm that is used to classify the given data by converging multiple base learners into strong learners. Bagging classifier also uses an ensemble approach to increase the predictive accuracy by training numerous base learners (e.g., DT) on different subsets of the train set and then combining their predictions to create a final classification. kNN[3] is a simple ML classifier that stores the training set and categorizes new data based on similarity. MLP is a model consisting of multiple layers and the nodes of the layers are neurons with nonlinear activation functions. There may be one or more nonlinear hidden layers between the input and the output layer.

[3]https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

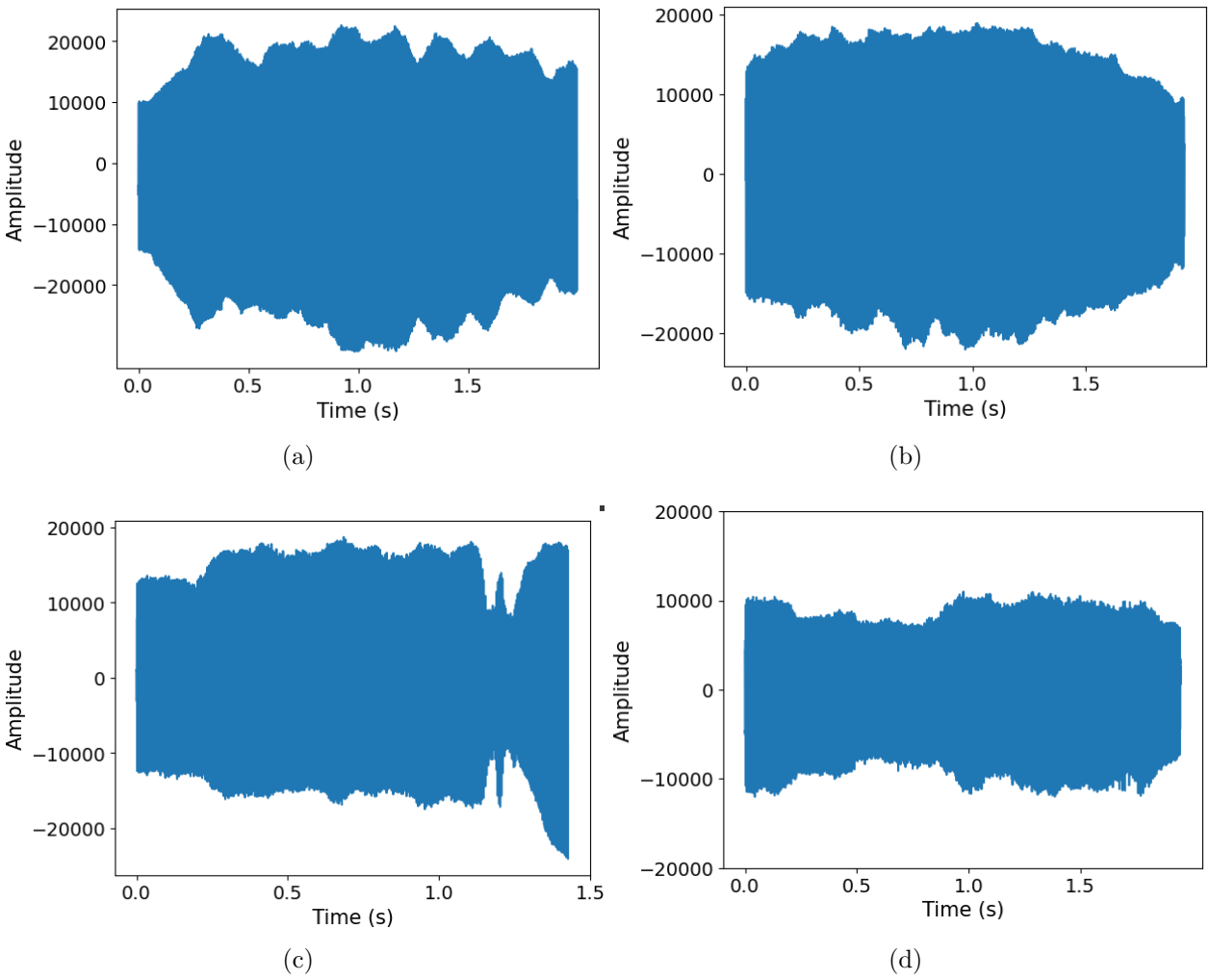


Figure 8: Waveforms of healthy voice samples at (a)high (b) low (c) low-high-low and (d) normal pitch

4.3 Experiments

Spectral features of the speech signals are extracted using Librosa[4] library and to balance the dataset, imbalanced-learn library[5] is used. To oversample the minority class (healthy class), the sampling strategy parameter is set to 'auto' while instantiating the oversampling algorithm (SMOTE) and the random state parameter which sets a seed value to control the randomness is set to 42, allowing consistent results which is important for comparisons. Table 11 gives the statistics of the dataset before and after applying the resampling techniques.

[4]https://librosa.org/doc/main/index.html
[5]https://imbalanced-learn.org/stable/

Table 11: Statistics of the dataset before and after balancing

	Imbalanced Data	Balanced Data
Healthy samples	868	1299
Pathological samples	1299	1299
Total	2167	2598

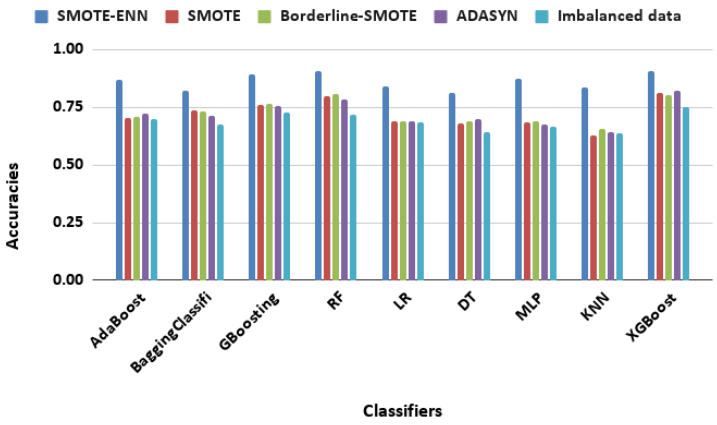


Figure 9: Comparison of the performances of ML models with and without resampling techniques

4.3.1 Dataset Description

In this work, speech signals of only sustained vowel /a/ with normal, low, low-high-low, and high pitches, for healthy and selected voice disorder (Psychogene Dysphonie, Hyperfunktionelle Dysphonie, Laryngitis, Reinke's Edema, Dysarthrophonie) samples available in SVD are used. Psychogene Dysphonie disorder (Psychogenic Dysphonia) causes hoarseness or difficulty in speaking without any observable physical or structural abnormalities in the vocal cords. Hyperfunktionelle Dysphonie disorder (Hyperfunctional Dysphonia) is the one in which the vocal apparatus is not used properly during speech production resulting in extreme muscle strain. Laryngitis disorder is characterised by swelling and inflammation of the vocal cords. Due to the vocal cords' reduced capacity to efficiently vibrate and produce sound, it frequently causes hoarseness or complete voice loss. Reinke's Edema is a particular kind of condition of the vocal cords that is characterized by swelling and fluid buildup in Reinke's space, a gelatinous layer of the vocal cords. Dysarthrophonie (Dysarthrophonia) is a voice disorder due to poor muscle control and coordination of the speech-producing muscles. Figure 8 illustrate waveforms of the speech signals of healthy samples at four different pitches.

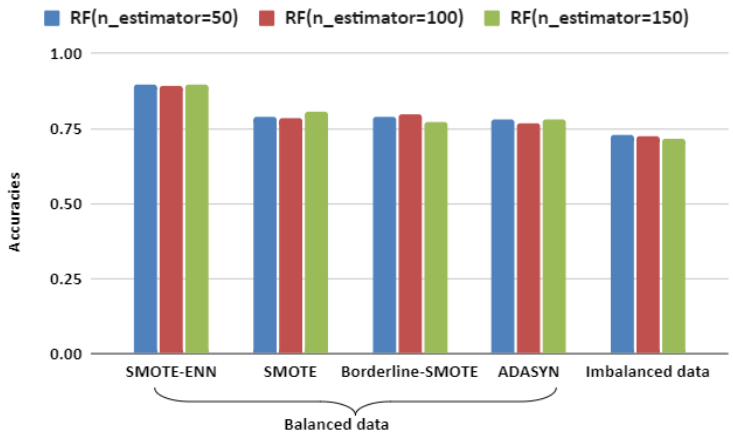


Figure 10: Comparison of the performances of fine-tuned RF models with and without resampling techniques

Table 12: Results of the XGBoost and RF models with resampling techniques

Data		Model	Accuracy	Precision	Recall	F1-score
Balanced data	SMOTE	XGBoost	0.810	0.831	0.789	0.809
		RF	0.804	0.829	0.777	0.802
	ADASYN	XGBoost	0.821	0.831	0.807	0.819
		RF	0.774	0.807	0.724	0.763
	Borderline-SMOTE	XGBoost	0.804	0.838	0.765	0.800
		RF	0.798	0.825	0.768	0.795
	SMOTE-ENN	XGBoost	**0.906**	**0.890**	**0.888**	**0.885**
		RF	**0.902**	**0.873**	**0.891**	**0.882**
Imbalanced data		XGBoost	0.750	0.773	0.824	0.798
		RF	0.726	0.740	0.836	0.785

4.4 Results and Discussion

The performances of the ML models with and without resampling techniques, measured in terms of accuracies are shown in Figure 9. The results clearly indicate that the accuracies of all the classifiers has not crossed 76% for imbalanced data as the models failed to classify the minority class samples, leading to a high number of false negatives. After applying the resampling techniques to balance the class distribution, the accuracies of the models have improved significantly, reaching upto 90%.

As RF and XGBoost models have performed better than other ML models, fine-tuning the hyperparameters is applied only for these two models to improve the performance further. n_estimator is a hyperparameter used in ensemble learning methods, like RF and XGBoost, to limit the number of DTs constructed by the ensemble model. Experiments are conducted by setting the n_estimators of RF model to 50, 100, and 150 and that of XGBoost model to

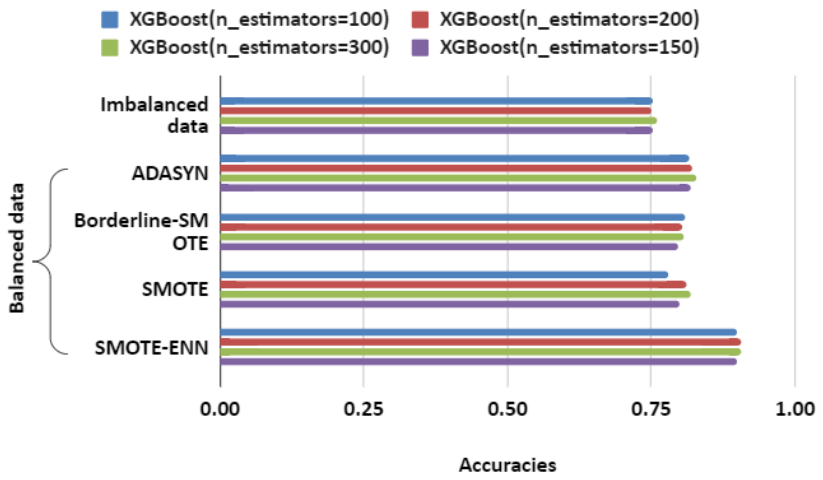


Figure 11: Comparison of the performances of fine-tuned XGBoost models with and without resampling techniques

100, 150, 200, and 300. The comparison of the performances of fine-tuned RF and XGBoost models with and without resampling techniques measured in terms of accuracies are shown in Figure 10 and 11, respectively. The results obtained by XGBoost and RF models, with and without resampling techniques (SMOTE, ADASYN, Borderline-SMOTE, and SMOTE-ENN) are shown in Table 12. The results clearly indicates that SMOTE-ENN has performed better for balancing the data, with accuracies of 0.902 and 0.906 for XGBoost and RF models respectively.

5 Proposed Work 4: Classification of Voice Disorder using Spectrograms of Voice Signals

In this work, we proposed a diagnostic system that can distinguish voice disorders precisely using spectrogram images of voice samples and using pre-trained models (VGG19, Inception-v3, ResNet50, VGG16, and DenseNet121) to extract deep features from spectrograms. Experiments are also conducted to enhance performance of the diagnostic system by concatenating features obtained from the two best pre-trained models. The proposed system is experimented on a subset of SVD that includes two classes: healthy and pathology. Results obtained demonstrate that the models using features extracted from i) ResNet50 and DensNet pre-trained models achieve better performance when used individually and ii) VGG16 and ResNet50 pre-trained models give better performance when the features are concatenated.

5.1 Related Work

Learning approaches have been used in many speech-related applications such as speaker recognition, speech emotion recognition, speech recognition, classification of voice disorders [54], etc. A comprehensive survey on voice disorder classification systems are found in [55][35]. Most of the researchers use SVD to detect voice disorders by extracting handcrafted features such as entropy, short-term cepstral parameters [47], Mel-frequency Cepstral Coefficients (MFCCs) [56], HNR, energy, and frequency. These features are used to train various ML [57][40] and DL [35] to classify voice disorders. Though various research works focus on using ML algorithms [57] [40] for voice disorder classification, the challenges such as manual feature extraction, model's ability to capture complex patterns from raw voice data especially captured in noisy environment needs to be addressed. On the other hand, DL models, such as CNN and LSTM, have shown enhanced performance by automatically extracting relevant features from the image data [58].

This has attracted the researchers to explore DL models for the classification of voice disorders using spectrogram representation. Chaiani et al. [54] applied CNN-LSTM to analyze spectrograms of disordered speech, achieving notable improvements in performance compared to ML methods. In another study, Huiyi Wu et al. [58] proposed a CNN-based approach for pathological voice detection using a subset of SVD by converting voice recordings of vowel /a/ to spectrograms. They initialized the weights in their model using Convolutional Deep Belief Network (CDBN) to make their proposed model more robust and prevent overfitting. Mittal et al. [59] developed voice disorder recognition framework by using stacked auto-encoder based DL approach for flexible decision-making, likely leading to more precise and robust detection results in voice pathology. They experimented on VOICED dataset comprising 150 pathological and 58 healthy voice recordings.

One of the major drawbacks of DL models is the requirement of large amount of labeled data for training, which is very challenging in medical field and this challenge is addressed using TL approach. Researchers have used image-based pre-trained models such as SqueezeNet1_1, ResNet101, DenseNet161 [60], ResNet34, AlexNet, ResNet-50, and Inception-v3 [36], to develop voice disorder detection system using spectrogram representation of voice signals. Due to deep architecture and extensive training on large-scale dataset like ImageNet, the image pre-trained models are well-known for their ability to capture robust visual representations.

5.2 Methodology

The proposed methodology consists of spectrogram creation by applying Fast Fourier Transform on voice signals followed by feature extraction using five pre-trained models (VGG19, Inception-v3, ResNet50, VGG16, and DenseNet121), and building the classification models to identify whether the given voice signal is healthy or disordered. The framework of

proposed methodology is shown in Figure 12. The background noise in the spectrograms are minimized by applying a low-pass filter and they are resized as per the requirements of pre-trained models.

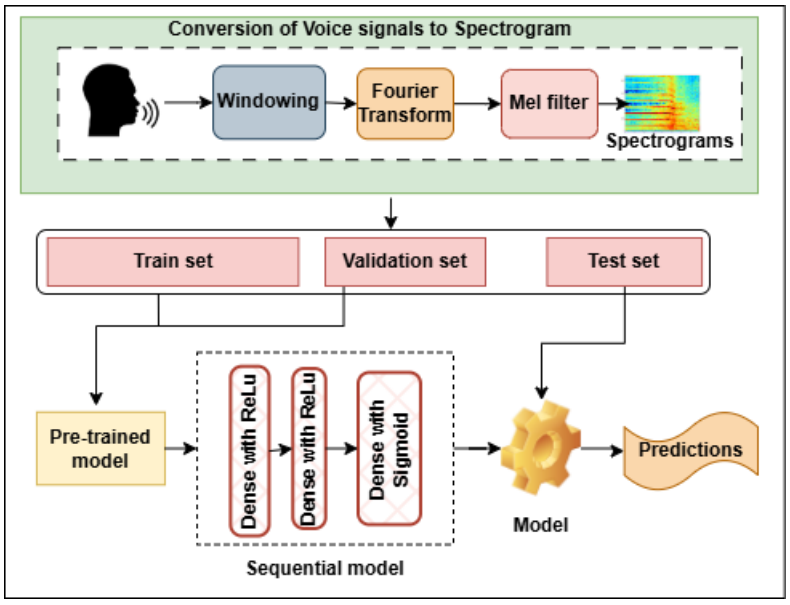
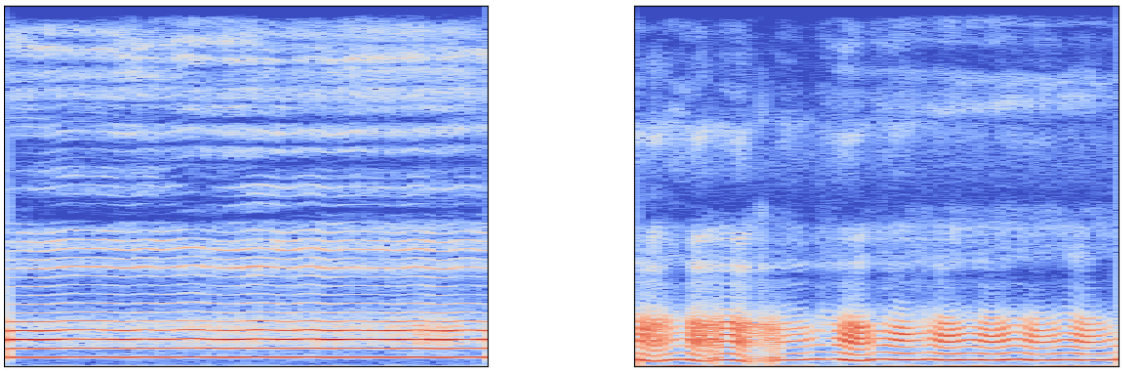


Figure 12: Framework of proposed TL approach



(a) (b)

Figure 13: Spectrogram representation of (a) Healthy sample and (b) Pathological sample

5.2.1 Feature Representation

Feature extraction is a method of extracting distinguishing features from the given data which are used to train the learning models. In this work, five image pre-trained models:

VGG19, Inception-v3, ResNet50, VGG16, and DenseNet121, are used for feature extraction from spectrograms.

The pre-trained models are loaded without the final fully connected layer as the focus is on extracting features from the earlier convolutional layers. The output of the last convolutional layer in the pre-trained model represents the high-level information of the input as features for subsequent processing.

5.2.2 Feature Concatenation

The features obtained by two different pre-trained models are concatenated to enhance the feature representation and framework of the proposed model with feature concatenation is shown in Figure 14. The deep features extracted from the pre-trained models are flattened to convert multidimensional feature maps into 1D vectors and the vectors from the two models are combined into a single vector that are utilized to train a sequential model. This is beneficial as it combines the features learned by two different architectures, potentially capturing different aspects of the input data, leading to better generalization and improved performance of the classification task.

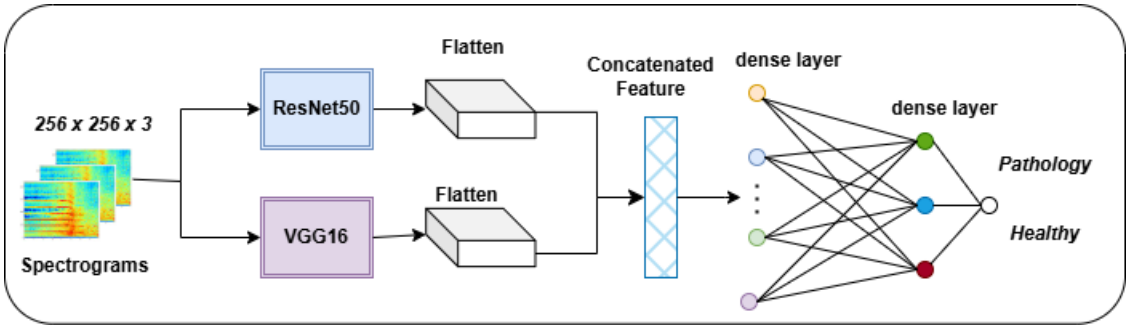


Figure 14: Framework of proposed model with feature concatenation

5.2.3 Classification Models

The classification model in the proposed method is a simple sequential model that consists of three dense layers. The features from the pre-trained model are flattened and then fed as input to the first dense layer (with 128 neurons), followed by the second dense layer (with 64 neurons). The first two dense layers uses ReLU activation function that helps the model learn complex patterns by introducing non-linearity. For binary classification, the final dense layer has a single neuron and it uses a sigmoid activation function to return a value between 0 and 1 that indicates the likelihood that the input falls into the pathological or healthy class.

Table 13: Hyper-parameters and their values

Hyper-Parameters	**Values**
Optimization Method	Adam
Activation function	ReLu, Sigmoid
Loss	binary_crossentropy
Batch Size	32
Total No. of Epochs	20

5.3 Experiments

Voice signals are converted to spectrogram images using Python libraries – Librosa and NumPy. As the pre-trained models accept the images of shape 256 x 256 with 3 color channels (RGB), the shape of the spectrogram images are set to 256 x 256 x 3.

5.3.1 Dataset Description

In the current study, only the vowel /a/ samples recorded at normal pitch are selected from SVD dataset. The vowel /a/ is commonly used in voice assessments because it's relatively easy to produce in a sustained manner, allowing for consistent measurements. The dataset is imbalanced with 1,354 pathology and 687 healthy samples. As the aim is to evaluate model performance on the original dataset, efforts are not made to balance the imbalance dataset. The samples of spectrograms is shown in Figure 13.

The dataset is divided into 1,633 samples for training and 204 samples each for Testing and Validation (Val). Various experiments are conducted with different hyperparameters and the best-performing configuration shown in Table 13 is employed in the proposed work. Adam optimizer is used for training the model due to its effective handling of sparse gradients and variable learning rate. The 'binary_crossentropy' loss function is ideal for binary classification problems as it is specially developed for models that use a sigmoid activation function in the last layer to output the probabilities. The model learns to create predictions that are closer to the true binary class by lowering binary cross entropy, which increases the overall performance and accuracy levels. Additionally, it focuses on minimizing errors between the predicted and actual outcomes for each binary decision. In order to avoid overfitting, early stopping is used, and the validation loss is monitored for five epochs. This configuration reduces the possibility of overfitting, maximizes training efficiency, and enhances generalization on the Val set.

Table 14: Performances of proposed models using pre-trained models for feature representation

Model	Val/Test set	AUC	Precision	Recall	Accuracy	Loss
VGG19	Val	99.76	97.40	99.34	97.69	0.0754
	Test	99.60	100	97.06	97.85	0.0726
ResNet50	Val	99.92	98.65	100	99.07	0.0364
	Test	**99.92**	**98.68**	**100**	**99.09**	**0.0376**
VGG16	Val	99.95	98.67	99.33	98.61	0.0329
	Test	99.50	98.14	100	98.63	0.0427
Inception-v3	Val	94.03	92.95	99.32	94.44	0.6701
	Test	95.81	95.41	98.94	96.09	0.8844
DenseNet121	Val	98.85	99.35	99.35	99.07	0.1138
	Test	**99.46**	**100**	**98.91**	**99.28**	**0.1290**

5.4 Results and Discussion

To quantitatively assess the performance of the proposed models, accuracy, precision, Area Under the Curve (AUC), loss, and recall, are adopted. Accuracy gives an overall success rate, while recall and precision help understand the model's performance in detecting true positives and minimizing false positives, respectively. AUC measures the overall performance of the binary classification model and the loss metric serves as an important indicator of how well the model predicts binary outcomes. During each epoch, the loss is calculated on both training and validation sets to assess the model's learning progress and its ability to generalize unseen data.

The performances of proposed models using the five pre-trained models for feature extraction are shown in Table 14. The results reveal that majority of the models have a loss in the range of 0.03 - 0.88 and 0.04 - 0.67 for the Test and Val set, respectively. Further, experiments are carried out by concatenating features extracted from two different pre-trained models and the results of five best combinations are shown in Table 15. From the result table, it is observed that all the models have obtained above 96% accuracy for both Val and Test sets. Further, feature concatenation has improved the results compared to considering features from single pre-trained models. The models trained using deep features derived from ResNet50 and VGG16 attained a high accuracy of 99.54% with a loss value of 0.0252. Similarly, feature concatenation of ResNet50 and DenseNet121 models obtained an accuracy of 99.28% with less loss value of 0.0150.

Among the pre-trained models used for feature extraction, Inception-v3 exhibited a higher loss value during validation compared to other models as shown in Table 14. ResNet50 and

Table 15: Performances of the proposed models using feature concatenation

Pretrained Models	Val/Test set	AUC	Precision	Recall	Accuracy	Loss
ResNet50 + VGG19	Val	99.86	96.95	99.37	97.22	0.0675
	Test	99.96	99.64	98.82	98.89	0.0290
ResNet50 + VGG16	Val	99.96	97.33	100	98.15	0.0313
	Test	**99.99**	**99.34**	**100**	**99.54**	**0.0252**
ResNet50 + Inception-v3	Val	96.33	95.45	100	96.76	0.2254
	Test	98.88	97.16	97.21	96.16	0.1327
ResNet50 + DenseNet121	Val	99.98	99.35	99.35	99.07	0.0251
	Test	**99.99**	**100**	**98.91**	**99.28**	**0.0150**
Inception-v3 + DenseNet121	Val	99.95	96.71	100	97.69	0.0636
	Test	99.65	100	99.32	99.54	0.0319

DenseNet121 pre-trained models demonstrated better generalization, as reflected by their lower validation loss and more stable loss progression, as shown in Table 15. Though the subset of SVD considered in this work is imbalanced and the dataset is small, the proposed models have obtained better performance due to the ability of pre-trained models that have already learned rich features from large-scale datasets.

6 Proposed Work 5: A Semi-supervised Co-training Strategy for Automated Voice Disorder Detection

Co-training is a semi-supervised learning technique where two models are trained simultaneously on different views or subsets of the data, allowing them to teach each other by sharing confident predictions [61]. In this work, co-training is applied to a KanVD dataset for the binary classification of voice disorders. The goal is to distinguish between healthy and disordered voices using limited labeled data and unlabeled samples. By incorporating high-confidence pseudo-labels generated during training, the model leverages unlabeled data to improve overall accuracy and generalization. This approach is particularly effective when labeled medical voice data is scarce.

6.1 Related Work

Co-training, first introduced by Blum and Mitchell [62] in 1998, is a popular semi-supervised learning strategy that relies on multiple classifiers trained on different views or subsets of

the data to iteratively label unlabeled samples [63]. Over the years, co-training has been successfully applied to various domains such as text classification [64], image classification [65], and medical diagnosis [66] [67]. In the field of speech and voice analysis, semi-supervised methods have gained attention due to the scarcity and high cost of annotated medical voice data. Recent studies have explored self-training and pseudo-labeling for speech-related tasks, but co-training remains relatively underexplored, especially in the context of voice disorder identification. However, its potential has been recognized in related areas such as emotion recognition [68] and gender detection from emotional speech [69], where co-training with spectral and temporal feature views has improved classification performance.

Zhao et al. [70] proposed an abnormal voice detection algorithm based on a semi-supervised co-training approach, specifically using an enhanced AR-Tri-training method. Their model combines multiple acoustic features such as MFCCs, spectral entropy, wavelet coefficients, and centroid measures. This work demonstrates the strength of co-training in leveraging unlabeled data for voice disorder detection, which aligns closely with the goals of our current study.

Given the limited availability of labeled pathological voice datasets, applying co-training for binary classification of voice disorders offers promising results. It allows the model to use large collections of unlabeled voice samples while maintaining robustness by using confidence-based sample selection.

6.2 Methodology

The goal of this study is to classify voice samples as either normal or pathological using semi-supervised approach. The dataset used in this experiment is subset of KanVD dataset and consisted exclusively of sustained vowel /a/ recordings, collected from both normal and pathological speakers. This study adopts a co-training framework for semi-supervised learning, combining a small set of labeled data with a larger set of unlabeled voice samples. The framework of binary classification is shown in Figure 15. The framework consists of the following key steps:

6.2.1 Data Preparation and Feature Extraction

In the initial phase of the co-training framework, the dataset is preprocessed and divided into two distinct subsets, (i) Labeled Data: The dataset containing voice samples with known class labels. This subset is used to train the initial classifier. (ii) Unlabeled Data: The remaining voice samples without labels are retained to be pseudo-labeled during the iterative co-training process.

Prior to the split, each voice recording is processed to extract MFCC using the librosa library. A total of 12 MFCCs are extracted per frame, and each sample is limited to a maximum of 50 frames. If a sample has fewer than 50 frames, zero-padding is applied;

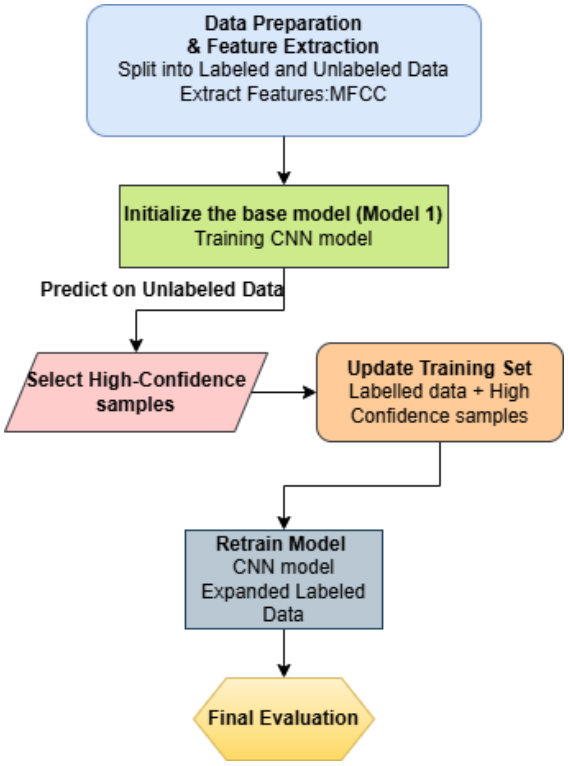


Figure 15: Co-training framework for binary classification of voice disorders

otherwise, it is truncated. The final MFCC shape for each sample is standardized to (20, 50, 1) to suit the input shape of a CNN.

6.2.2 Model Architecture

A CNN is designed to classify the MFCC features. The architecture included two convolutional layers followed by max-pooling, flattening, and two fully connected layers, ending with a softmax output layer for binary classification (normal vs. pathological). The model is trained using the Adam optimizer and sparse categorical crossentropy loss.

6.2.3 Semi-supervised Learning using Pseudo-Labeling

To improve performance in the presence of limited labeled data, a semi-supervised learning approach is implemented using pseudo-labeling. The model is first trained on a labeled subset of the dataset (70%). Predictions are then made on the remaining 30% unlabeled data, and samples with high-confidence predictions are selected as pseudo-labeled data. The following four confidence thresholds are explored:

- Threshold 1: 0.80
- Threshold 2: 0.85

Table 16: Overview of Voice Dataset Utilized in Co-Training

Dataset Description	
Dataset	Subset of KanVD + Private data
Classes	Healthy / Pathology
Recordings	Sustained vowel /a/
Total samples	972
Labeled data	680
Samples	Healthy - 217 Pathology -463
Unlabeled data	292

- Threshold 3: 0.90
- Threshold 4: 0.95

High-confidence samples (where the maximum softmax probability exceeded the threshold) are added to the labeled dataset, and the model is fine-tuned on the expanded dataset.

6.3 Experiments

6.3.1 Dataset Description

The dataset used in this study comprises voice recordings collected for the purpose of pathological voice disorder detection. It includes both healthy and disordered samples, with disorders such as vocal fold nodules, polyps, and paralysis. Each audio sample is processed to extract MFCC features, ensuring uniformity in input representation. The dataset is divided into labeled and unlabeled subsets to facilitate the co-training process. Table 16 describes the information of the dataset used in the experiments.

As part of the experimental process, a confidence-based pseudo-labeling approach to improve model performance through semi-supervised learning is implemented. In this method, the first step involved obtaining the predicted probabilities for the unlabeled data using the trained model. Next, pseudo-labels are assigned to each sample based on the class with the highest predicted probability. To assess the reliability of these predictions, the maximum probability for each sample is extracted and treated as the model's confidence level. Finally, to ensure the quality of the pseudo-labeled data, only those samples with high confidence-specifically, those with a confidence score greater than confidence thresholds are selected. These high-confidence pseudo-labeled samples are then used for further training, with the aim of enhancing the model's generalization.

The dataset is split into training and testing sets in a 70:30 ratio using stratified sampling to maintain class balance. The initial CNN model is trained using only the labeled data for 25 epochs, with 30% of the training data used for validation. Subsequently, pseudo-labels

Table 17: Classification performance of CNN model with pseudo-labeling at different confidence thresholds

Confidence Threshold	**Confidence samples**	**Accuracy**	**Loss**
No pseudo-labeling	without conf-threshold	80.15	0.8630
> 0.80	**262**	**94.18**	**0.1506**
> 0.85	260	94.15	0.1362
> 0.90	243	93.51	0.1874
> 0.95	221	92.27	0.2257

are generated from the unlabeled test set, and the model is re-trained for an additional 20 epochs with the augmented dataset. For evaluation, Accuracy and Loss functions are used which computed on the unseen test set before and after pseudo-labeling to observe any performance improvements.

6.4 Results and Discussion

The classification performance of the CNN model with different pseudo-labeling confidence thresholds is shown in Table 17. The best performance is observed when the pseudo-label confidence threshold is set to 0.80, indicating that incorporating confidently predicted samples significantly enhanced the model's generalization capabilities. However, as the threshold increased, fewer samples are selected for augmentation, resulting in a smaller performance gain.

During the experiments, the lower confidence threshold, such as > 0.80, resulted in a larger number of pseudo-labeled samples being added to the training set. Although some of these samples may have been mislabeled, the increased volume of data helped the model generalize better, leading to improved overall performance. This is because the model is exposed to more diverse examples, which contributed to better learning despite the presence of some noise. On the other hand, when a higher confidence threshold, such as > 0.95, is used, the pseudo-labeled samples are much cleaner but significantly fewer in number. This limited the variety of training data, which in some cases caused the model to underfit due to insufficient exposure to diverse patterns. Since early stopping is applied, the model trained for only a few epochs when using less data, as it quickly met the stopping criteria due to limited training samples. Thus, there is a trade-off between label quality and data quantity that impacted model performance.

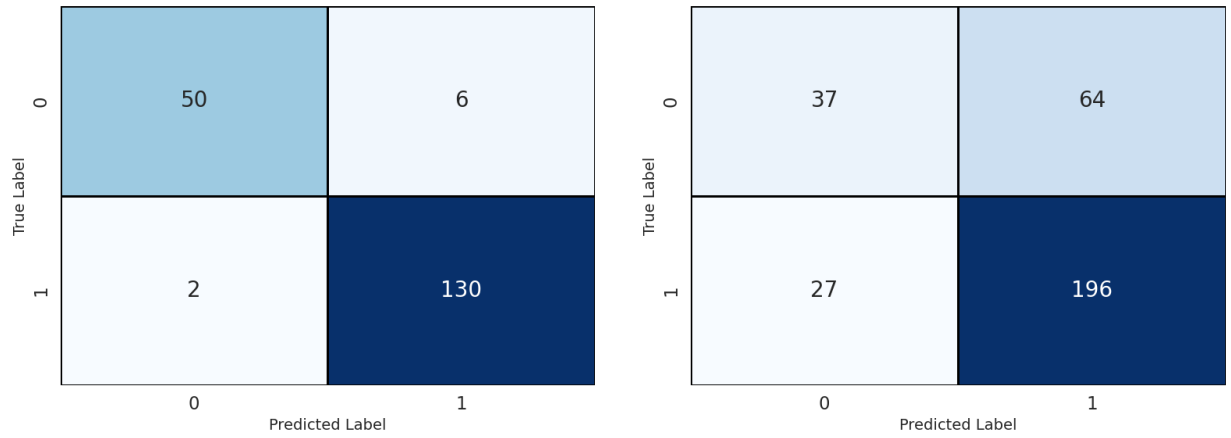
6.5 Comparative Study

To validate the robustness and generalizability of our approach, we compared the results with those obtained from the SVD dataset. Our model demonstrated competitive performance, despite differences in recording conditions, language, and speaker diversity present in the

Table 18: Classification performance of CNN model using on SVD and KanVD dataset

Dataset	No. Pseudo-labeling	>0.80	>0.85	>0.90	>0.95
KanVD	Confidence samples	**262**	260	243	221
	Accuracy	**94.18**	94.15	93.51	92.27
SVD	Confidence samples	**227**	189	158	106
	Accuracy	**93.35**	70.99	73.90	69.06

SVD dataset. Furthermore, the use of distinct pseudo-labeling strategies in the co-learning setup significantly contributed to improved classification accuracy by leveraging unlabeled data effectively. Table 18 presents a comparative evaluation of the proposed methodology when trained on KanVD dataset versus the SVD dataset, showing a consistent outperformance on our dataset (e.g., 0.83% higher accuracy). This highlights the model's improved ability to capture acoustic features inherent to our data. The confusion matrix for the best-performing model is shown in Figure 16.



(a) (b)

Figure 16: Confusion matrix showing the performance of the model on the >85 Confidence Threshold on (a) KanVD dataset and (b) SVD dataset

7 Proposed Work 6: Automatic Assessment of Pathological Voice Quality using Learning Approaches

The objective of this study is to identify the voice qualities of the given unlabeled sustained voice samples and the problem is modeled as a multi-class classification problem with four labels - Hoarse, Harsh, Breathy, and Healthy.

7.1 Dataset Description

Voice samples are obtained from KanVD dataset. Kannada-speaking individuals are chosen to understand the impact of different voice qualities in that population. The voice samples of sustained phonation of /a/ are recorded using the MDVP module of CSL with a sampling frequency of 44.1kHz. Additionally, voice samples are also collected from people in the different age group with healthy voices following the same procedure. The selection of people having healthy voice is based upon a clear, normal-sounding voice in conversational speech as judged by experienced SLPs. Three experienced SLPs who have diagnosed and treated voice issues for more than three years participated in the subjective assessment for perceptual parameters. Samples are classified as hoarse, harsh, breathy, and healthy by SLPs according to the factors of the GRBAS scale (degree of hoarseness, roughness, breathiness).

A total number of 428 sustained phonation samples in the age group 18 to 75 were collected in clinical settings. For voice pathologies with limited cases, multiple samples are obtained from the individuals in order to have a comparatively good number of samples. It also helps to capture variations within a participant's voice patterns, which is important for developing robust models. The distribution of voice samples with respect to voice quality is shown in Figure 17. As the number of voice samples varies significantly across the voice qualities, the dataset is imbalanced. The primary purpose is to gain insights directly from the actual dataset in its original form. So, to retain the inherent characteristics of data and identify the patterns obtained from the data, the data is not balanced.

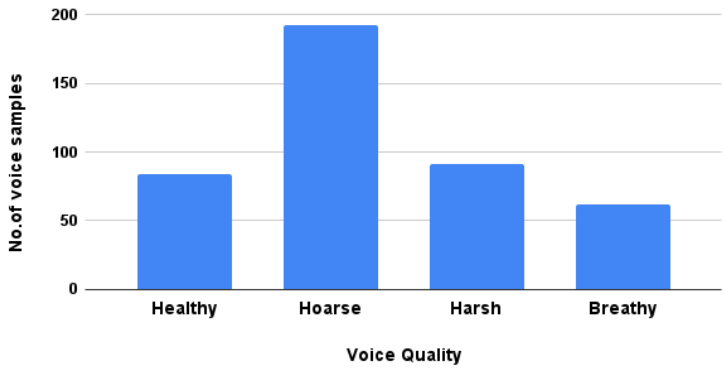


Figure 17: Voice sample distribution with respect to voice quality

7.2 Related Work

Voice signal is used in several applications ranging from emotion recognition [71] to patient healthcare [72]. Several non-invasive solutions, including identification of voice disorders [73] [30] [56] and identification of the severity of voice [74], have used voice signals to estimate the state of voice health. Over recent years, several ML models have been developed to improve

the performance of voice disorder detection but identification of voice quality [75] is in its infancy.

Al-Nasheri et al. [73], explored detection and classification of voice disorders based on the analysis of different frequency bands using auto-correlation and entropy. They used sustained vowel /a/ samples of both normal and disordered voices from MEEI, SVD, and AVPD datasets to train SVM and obtained accuracies of 99.54%, 99.53%, and 96.02% for MEEI, SVD, and AVPD, respectively. SVM technique is also used by Souissi et al. [29] to estimate the presence of four types of Dysphonia: Chronic laryngitis, Cysts, Reinke's edema, and Spasmodic dysphonia, considering 70 pathological and 40 healthy voices from the SVD dataset. Using MFCC features by ignoring their first and second derivative and LDA for dimensionality reduction method, they obtained an accuracy of 86% for voice disorders identification. Verde et al. [47] performed a comparative study of different classifiers (SVM, DT, Bayesian Classification, Logistic Model Tree (LMT), and Instance-based Learning algorithms) to identify voice pathology using handcrafted features (F0, jitter, shimmer, HNR, MFCC (from 1 to 13), and first derivative and second derivative) extracted from the sustained vowel /a/ samples collected from SVD dataset.

In summary, related works illustrate that many ML algorithms are explored for the classification of voice disorders using publically available datasets: SVD, AVPD, and MEEI [51]. But, very less work is reported on voice quality identification due to lack of voice quality annotated data which is very important in identifying voice qualities. This gives scope to explore different ML algorithms to identify voice qualities of given unlabeled sustained phonation samples using various acoustic features.

7.3 Methodology

The voice quality identification system is aimed to identify the voice qualities of given unlabeled sustained voice samples. Figure 18 depicts workflow of the proposed model and steps involved in building the proposed models are explained below:

7.3.1 Pre-processing and Feature Extraction

Data pre-processing plays an important role in improving the quality and relevance of the data. The voice samples in the dataset are in .nsp format and they are converted to .wav format using GoldWave software[6], for further processing. In order to extract the acoustic features, the sample rate of voice samples is adjusted to 22.05 kHz as smaller sample rates can result in faster processing and smaller computational loads since they require processing of fewer data points.

Acoustic features offer valuable insights into speech patterns and vocal characteristics

[6]https://www.goldwave.com/

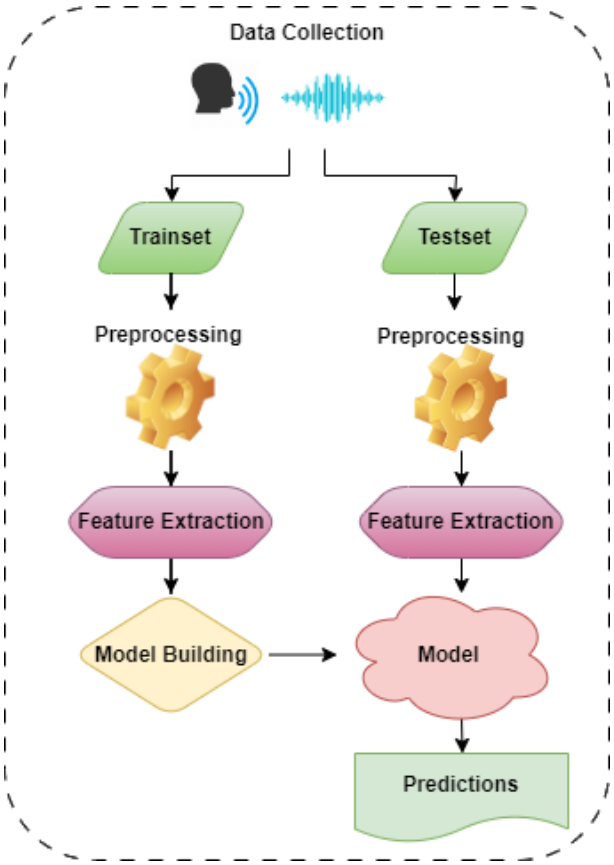


Figure 18: Workflow of the proposed voice quality classification system

and hence the acoustic features (MFCCs, Spectral centroid, Spectral roll-off, ZCR, RMS, Chroma stft, Spectral contrast, and Mel spectrogram) are extracted from the given voice samples. These features are extracted using Librosa library and the number of features obtained from each feature type is shown in Table 19. As each feature provides a unique characteristic, combining the features aids in improved performance to identify the voice quality. Hence, the individual features are combined to form the five feature sets as shown in Table 20 and these feature sets are used to train the ML classifiers individually.

7.3.2 Model Construction

With sufficient training data, ML models learn the relationship between features and the corresponding labels. In the proposed work, ML classifiers (Bagging Classifier, ExtraTrees Classifier, GB, XGB, RF, LR, and DT) are used. Table 21 shows the parameters and their default values of ML models used in the experiments.

Table 19: Acoustic features and their values

Features	No. of Features
Mel spectrogram	128
MFCCs	13
Chroma stft	12
Spectral contrast	7
Spectral centroid	1
ZCR	1
Spectral rolloff	1
RMS	1

Table 20: Combination of features

Feature set	Features combination
F_set1	MFCCs
F_set2	Mel spectrogram
F_set3	Spectral contrast + Mel spectrogram
F_set4	F_set3 + Spectral centroid + Spectral roll-off + ZCR + RMS + Chroma stft
F_set5	F_set1 + F_set4

7.4 Experimental Results and Discussion

Pandas library is used for data pre-processing, Librosa library for loading and feature extraction, and scikit-learn is used to build the classifiers. Based on the preliminary understanding of relevant features typically used in previous studies [30] [56], features are combined and systematically tested to assess their contribution to the model's performance. These features were selected for their ability to capture subtle variations in voice quality, such as harshness, hoarseness, breathiness, and normal phonation, which signify underlying voice pathologies. In this study, various experiments are conducted for the identification of voice quality and the performances of the proposed ML models are measured in terms of precision, recall, F1-score, and accuracy. For evaluating the effectiveness of ML models, the dataset is split into train and test sets in the ratios - 60:40, 70:30, and 80:20, and models are trained on the feature sets shown in Table 20.

The performances of the proposed ML models on 60:40, 70:30, and 80:20 train-test set ratios are shown in Tables 22 and 23, respectively and the best results are highlighted. Among the experiments conducted in three different train-test set ratios, all the classifiers in 80:20 have performed better than the other two train-test set ratios. Further, in 80:20 train-test ratio, the ExtraTrees classifier trained with feature set F_set3 (Spectral contrast and Mel spectrogram) achieved the top results of 89.54% of accuracy, 90.14% of precision, and

Table 21: Parameters of ML models and their values

Classifier	Parameters
BC	n_estimators=10, random_state=None
GB	loss = 'log_loss', learning_rate=0.1)
XGB	learning_rate = 0.1, n_estimators = 100
RF	criterion = 'gini', n_estimators = 100
DT	splitter='best', max_depth=2, random_state=42
LR	solver='lbfgs', multi_class='multinomial'
ETC	n_estimators = 100, criterion = 'gini'

Table 22: Performances of ML models in 70:30 and 60-40 train-test set ratio

Feature set	Train-test split	BC	ETC	GB	XGB	RF	LR	DT
F_set1	70-30	70.54	79.85	72.87	75.19	82.17	55.81	65.12
	60-40	66.28	79.65	72.67	73.84	77.91	51.74	61.63
F_set2	70-30	74.42	81.40	75.97	80.62	75.19	60.47	61.24
	60-40	63.37	77.33	70.35	75.00	72.67	54.65	62.79
F_set3	70-30	75.97	82.95	82.17	78.30	82.95	61.24	70.54
	60-40	62.79	81.40	78.49	73.84	75.00	55.81	56.98
F_set4	70-30	72.87	86.05	84.50	80.62	80.62	47.29	61.24
	60-40	66.86	83.72	75.00	71.51	74.42	44.77	59.30

86.84% of recall. This illustrates that combination of Spectral contrast and Mel spectrogram features are effective in capturing the relevant characteristics of the voice quality.

Among the evaluated models, Extra Trees and XGBoost consistently outperformed others, such as SVM and Random Forest, achieving up to 89.54% accuracy and 0.8807 F1-score on multi-classification tasks. Extra Trees excelled due to its randomized node splitting and deeper ensemble averaging, which mitigated overfitting in high-dimensional acoustic features while enhancing generalization on noisy voice data. XGBoost's gradient boosting framework, with built-in regularization and efficient handling of class imbalances in pathological samples, provided superior feature interaction modeling.

This work has exhibited that even a small dataset would be acceptable to train the proposed ML models to yield better performance. However, the voice quality dataset exhibits an imbalanced distribution of samples across the four classes having significantly more 'hoarse' samples than others. This imbalance may result in the model being biased towards the majority class (i.e. hoarse), leading to an increased chance of false positives for that class. Hence, this imbalance must be addressed during training to ensure fair representation and

Table 23: Performances of ML models in 80:20 train-test set ratio

Feature set	Evaluation metrics	BC	ETC	GB	XGB	RF	LR	DT
F_set1	**Accuracy**	77.91	82.56	77.91	72.09	79.07	55.81	68.61
	Precision	73.38	79.94	71.42	66.80	80.33	45.67	60.97
	Recall	68.37	76.80	68.94	68.09	73.58	43.18	70.08
	F1-score	70.17	78.01	69.88	66.49	76.03	43.22	63.11
F_set2	**Accuracy**	70.93	86.05	80.23	**86.05**	82.56	62.79	74.42
	Precision	69.39	88.14	77.42	**80.62**	86.98	63.73	67.28
	Recall	66.76	80.02	74.72	**82.01**	73.30	54.83	70.27
	F1-score	65.93	83.07	75.59	**81.21**	77.81	55.32	66.78
F_set3	**Accuracy**	76.74	**89.54**	84.88	83.72	88.37	61.63	67.44
	Precision	73.20	**90.14**	81.11	77.71	89.18	55.14	60.40
	Recall	74.05	**86.84**	77.84	80.87	81.16	58.14	67.99
	F1-score	72.71	**88.07**	78.74	78.87	84.16	55.40	62.27
F_set4	**Accuracy**	76.74	88.37	81.40	81.40	84.88	50.00	68.61
	Precision	71.10	90.18	76.32	78.23	88.28	45.50	66.28
	Recall	75.57	81.16	77.75	79.73	74.34	34.47	73.67
	F1-score	72.81	84.66	76.93	78.15	78.97	34.02	66.49
F_set5	**Accuracy**	76.74	86.05	87.21	84.88	87.21	59.30	69.77
	Precision	74.69	88.68	85.19	79.48	89.67	53.45	60.80
	Recall	75.00	78.50	84.19	79.92	80.59	51.04	66.57
	F1-score	74.30	82.53	84.63	79.66	84.13	49.58	61.51

prevent bias.

8 Conclusion and Future Works

In this study, we focused on the classification of Dysphonia using both ML and DL techniques. A key part of the work involved the creation of a custom voice disorder dataset named KanVD, specifically collected from the Kannada-speaking population. Initially, efforts were dedicated to the development of this dataset, followed by establishing benchmark results using baseline models. To improve performance, we experimented with various acoustic feature sets and explored multiple ML algorithms and DL architectures. In the later phase of the study, recognizing the availability of additional unannotated voice samples, we implemented a semi-supervised learning approach using co-training to use the unlabeled data effectively. Furthermore, experiments on the SVD dataset employed two primary approaches: spectrogram generation followed by feature extraction via a pre-trained model and classification, and oversampling approach to address data imbalance and scarcity. The developed dataset was also compared against a publicly available voice disorder dataset (i.e., SVD), and the results demonstrated that our dataset offered more reliable and contextually appropriate classification performance, particularly for Indian speakers.

Future Work will focus on the following directions:

- Expanding the dataset with more samples across different regions in India.
- Validating the system with clinical support from ENT specialists and SLPs.
- Developing a software tool or mobile application using KanVD dataset, tailored for Indian hospitals, speech therapy centers, and telemedicine platforms.
- Collaborating with healthcare institutions to build a real-time voice disorder screening system for rural and urban populations.

By advancing in these directions, the research aims to contribute towards accessible and accurate voice disorder detection systems in India, especially in Karnataka.

9 Publications

9.1 Publications in International Conferences

- Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. "**Identification of Voice disorders: A Comparative Study of Machine Learning Algorithms.**" *In International Conference on Speech and Computer*, pp. 565-578. Cham: Springer Nature Switzerland, 2023.

- Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, Shwetha Prabhu, and Hemaraja Nayaka S. "**Automatic Assessment of Pathological Voice Quality using Machine Learning Approaches**" *In Third edition of the International Conference on Speech & Language Technology for Low-resource Languages (SPELLL).* Cham: Springer, 2024.

- Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, Shwetha Prabhu, and Hemaraja Nayaka S. "**Classification of Voice Disorders using Spectrograms of Voice Signals**" *In International Conference on International Conference on Artificial Intelligence in Healthcare.* Cham: Springer, 2025. (Received Best Research Paper Award)

- Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Shwetha Prabhu, "**KanVD: Corpus Creation of Kannada Voice Disorder Dataset**",*IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (2025 IEEE DISCOVER)* (Paper Accepted)

9.2 Papers Presented in National Conferences

- Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, Shwetha Prabhu, and Cynthia Santhmayor. "**Automated Voice Disorder Identification using Learning Approaches**"

- Sharal Coelho, Swarnamanaswini, Hosahalli Lakshmaiah Shashirekha, and Shwetha Prabhu. "**Enhanced voice disorder classification using pre-trained models**" (Received Best Paper Award).

Both papers were presented at "***FOCUS-4: Artificial Intelligence-Revolutionizing Communication And Hearing Healthcare***" a National-level Conference conducted by Father Muller College (Speech and Hearing), Mangalore.

9.3 Papers Under Preparation

- A Semi-supervised Co-training Strategy for Automated Voice Disorder Detection

10 Organization of the Thesis

The thesis consists of eight chapters along with appendices and is organized as follows:

- Chapter 1 - includes introduction of Voice disorders, Causes, and AI based detection techniques, overview of Machine Learning framework, publicly available datasets for diagnostic study, limitations of existing methods, motivation and objectives followed by organization of the thesis.

- Chapter 2 - presents construction of KanVD datasets for Dysphonia classification along with baselines.
- Chapter 3 - introduces a classification framework for Dysphonia/voice disorder, using acoustic feature sets and machine learning algorithms.
- Chapter 4 - presents classification of voice disorders using resampling techniques (SMOTE, ADASYN, Borderline-SMOTE, and SMOTE-ENN).
- Chapter 5 - presents work classification of voice disorders using spectrogram obtained from speech signals.
- Chapter 6 - presents a semi-supervised co-training strategy for automated voice disorder detection.
- Chapter 7 – presents automatic assessment of Dysphonia voice quality using learning approaches.
- Chapter 8 - presents the conclusion and future works.
- Appendices: Memorandum of Understanding (MoU), Consent forms, Ethical clearance certificates, and related materials.

References

[1] G'ulomjonova Shoira and IA Egamberdiyeva. Speech comunication is the tool. *Modern education and development*, 25(4):114–120, 2025.

[2] Soo Jin Park, Caroline Sigouin, Jody Kreiman, Patricia A Keating, Jinxi Guo, Gary Yeung, Fang-Yu Kuo, and Abeer Alwan. Speaker identity and voice quality: Modeling human responses and automatic speaker recognition. In *Interspeech*, pages 1044–1048, 2016.

[3] Douglas O'Shaughnessy. Spoken language identification: An overview of past and present research trends. *Speech Communication*, 167:103167, 2025.

[4] Saeid Safavi, Martin Russell, and Peter Jančovič. Automatic speaker, age-group and gender identification from children™s speech. *Computer Speech & Language*, 50:141–156, 2018.

[5] Janina K Casper and Rebecca Leonard. *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Lippincott Williams & Wilkins, 2006.

[6] Diane W Chen and Julina Ongkasuwan. Spasmodic dysphonia. *International ophthalmology clinics*, 58(1):77–87, 2018.

[7] Jan Willem Dankbaar and Frank A Pameijer. Vocal cord paralysis: anatomy, imaging and pathology. *Insights into imaging*, 5(6):743–751, 2014.

[8] Paulo Eduardo Przysiezny and Luciana Tironi Sanson Przysiezny. Work-related voice disorder. *Brazilian Journal of Otorhinolaryngology*, 81(2):202–211, 2015.

[9] Yu Zhang and Jack J Jiang. Acoustic Analyses of Sustained and Running Voices from Patients with Laryngeal Pathologies. *Journal of Voice*, 22(1):1–9, 2008.

[10] Mohammed Zakariah, Yousef Ajmi Alotaibi, Yanhui Guo, Kiet Tran-Trung, Mohammad Mamun Elahi, et al. An Analytical Study of Speech Pathology Detection Based on MFCC and Deep Neural Networks. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[11] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia technology*, 9:1112–1122, 2013.

[12] João Paulo Teixeira and Paula Odete Fernandes. Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices. *Procedia technology*, 16:1228–1237, 2014.

[13] Rumana Islam and Mohammed Tarique. Spectrogram and mel-spectrogram based dysphonic voice detection using convolutional neural network. In *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET*, pages 1–5. IEEE, 2024.

[14] Vibhuti Gupta. Voice Disorder Detection using Long Short Term Memory (LSTM) Model. *arXiv preprint arXiv:1812.01779*, 2018.

[15] Jia Min Karen Kua, Tharmarajah Thiruvaran, Mohaddeseh Nosratighods, Eliathamby Ambikairajah, and Julien Epps. Investigation of spectral centroid magnitude and frequency for speaker recognition. In *Odyssey*, page 7, 2010.

[16] Rajesh G Bachu, S Kopparthi, B Adapa, and Buket D Barkana. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In *Advanced techniques in computing sciences and software engineering*, pages 279–282. Springer, 2009.

[17] Marko Kos, Zdravko Kačič, and Damjan Vlaj. Acoustic classification and segmentation using modified spectral roll-off and variance-based features. *Digital Signal Processing*, 23(2):659–674, 2013.

[18] Haewon Byeon and Sunghyoun Cho. The factors of subjective voice disorder using integrated method of decision tree and multi-layer perceptron artificial neural network algorithm. *International Journal of Advanced Computer Science and Applications*, 7(5), 2016.

[19] GB Gour, V Udayashankara, Dinesh K Badakh, and Yogesh A Kulkarni. Voice-Disorder Identification of Laryngeal Cancer Patients. *International Journal of Advanced Computer Science and Applications*, 11(11), 2020.

[20] Maher Maalouf. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 3(3):281–299, 2011.

[21] Shintaro Fujimura, Tsuyoshi Kojima, Yusuke Okanoue, Kazuhiko Shoji, Masato Inoue, Koichi Omori, and Ryusuke Hori. Classification of voice disorders using a one-dimensional convolutional neural network. *Journal of Voice*, 36(1):15–20, 2022.

[22] Xiangyu Peng, Huoyao Xu, Jie Liu, Junlang Wang, and Chaoming He. Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports*, 13(1):7264, 2023.

[23] Alisson Constantine-Macías, Carlos Jiménez-Farfán, Michael Arce, Josué Tomalá, Edwin Valarezo-Añazco, Enrique Peláez Jarrín, Juan Solá, Carmen Gasca-Salas, Jorge U Máñez-Miró, Clara Trompeta, et al. Convolutional neural network-based classification

of parkinsonian voice disorders: A spectrogram approach. In *2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM)*, pages 1–6. IEEE, 2024.

[24] José Alberto Souza Paulino, Herman Martins Gomes, Leonardo Vidal Batista, and Leonardo Wanderley Lopes. Analysis of frequency range effect on the detection of voice disorder using convolutional neural networks trained on spectogram images. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 01–06. IEEE, 2024.

[25] Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9):4419, 2022.

[26] Florin Bogdan and Mihaela-Ruxandra Lascu. Advances and challenges in deep learning for acoustic pathology detection: A review. *Technologies*, 13(8):329, 2025.

[27] Ahmed Al-Nasheri, Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, and Mohamed A Bencherif. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, 31(1):113–e9, 2017.

[28] Ugo Cesari, Giuseppe De Pietro, Elio Marciano, Ciro Niri, Giovanna Sannino, and Laura Verde. A new database of healthy and pathological voices. *Computers & Electrical Engineering*, 68:310–321, 2018.

[29] Nawel Souissi and Adnane Cherif. Dimensionality Reduction for Voice Disorders Identification System Based on Mel Frequency Cepstral Coefficients and Support Vector Machine. In *2015 7th international conference on modelling, identification and control (ICMIC)*, pages 1–6. IEEE, 2015.

[30] Sharal Coelho and Hosahalli Lakshmaiah Shashirekha. Identification of voice disorders: a comparative study of machine learning algorithms. In *International Conference on Speech and Computer*, pages 565–578. Springer, 2023.

[31] Nuha Qais Abdulmajeed, Belal Al-Khateeb, and Mazin Abed Mohammed. A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions. *Journal of Intelligent Systems*, 31(1):855–875, 2022.

[32] Mark Huckvale, Zhuoya Liu, and Catinca Buciuleac. Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech. *Biomedical Signal Processing and Control*, 86:105201, 2023.

[33] Katia Nemr, Marcia Simoes-Zenari, Gislaine Ferro Cordeiro, Domingos Tsuji, Allex Itar Ogawa, Maysa Tibério Ubrig, and Márcia Helena Moreira Menezes. Grbas and cape-v scales: high reliability and consensus when applied at different times. *Journal of voice*, 26(6):812–e17, 2012.

[34] James Thomas and Sarah McNabb. Principles and ethics of collecting and managing health data. In *The Palgrave Handbook of Global Health Data Methods for Policy and Practice*, pages 469–486. Springer, 2019.

[35] Sarika Hegde, Surendra Shetty, Smitha Rai, and Thejaswi Dodderi. A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6):947–e11, 2019.

[36] Roohum Jegan and R Jayagowri. Enhancing Voice Disorder Detection Using Deep Transfer Learning Feature Fusion. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, volume 2, pages 1–6. IEEE, 2024.

[37] Christopher L Payten, Greg Chiapello, Kelly A Weir, and Catherine J Madill. Frameworks, terminology and definitions used for the classification of voice disorders: a scoping review. *Journal of Voice*, 38(5):1070–1087, 2024.

[38] Vyom Verma, Anish Benjwal, Amit Chhabra, Sunil K Singh, Sudhakar Kumar, Brij B Gupta, Varsha Arya, and Kwok Tai Chui. A novel hybrid model integrating mfcc and acoustic parameters for voice disorder detection. *Scientific Reports*, 13(1):22719, 2023.

[39] Fahad Taha AL-Dhief, Nurul Mu'azzah Abdul Latiff, Marina Mat Baki, Nik Noordini Nik Abd Malik, Naseer Sabri, and Musatafa Abbas Abbood Albadr. Voice pathology detection using support vector machine based on different number of voice signals. In *2021 26th IEEE Asia-Pacific Conference on Communications (APCC)*, pages 1–6. IEEE, 2021.

[40] Manjit Singh Sidhu, Nur Atiqah Abdul Latib, and Kirandeep Kaur Sidhu. MFCC in Audio Signal Processing for Voice Disorder: A Review. *Multimedia Tools and Applications*, pages 1–21, 2024.

[41] Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, Mansour Alsulaiman, Zulfiqar Ali, Ahmed Al-Nasheri, and Ghulam Muhammad. Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms. *Journal of healthcare engineering*, 2017(1):8783751, 2017.

[42] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. A Study of using Cough Sounds and Deep Neural Networks for the Early Detection of COVID-19. *Biomedical Engineering Advances*, 3:100025, 2022.

[43] Ji-Na Lee and Ji-Yeoun Lee. An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection. *Applied Sciences*, 13(6):3571, 2023.

[44] Yuanbo Wu, Changwei Zhou, Ziqi Fan, Di Wu, Xiaojun Zhang, and Zhi Tao. Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection. *IEEE Access*, 9:30–44, 2020.

[45] Mohamed Dahmani and Mhania Guerti. Glottal Signal Parameters as Features Set for Neurological Voice Disorders Diagnosis using K-Nearest Neighbors (KNN). In *2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–5. IEEE, 2018.

[46] Asli Nur Omeroglu, Hussein MA Mohammed, and Emin Argun Oral. Multi-modal Voice Pathology Detection Architecture based on Deep and Handcrafted Feature Fusion. *Engineering Science and Technology, an International Journal*, 36:101148, 2022.

[47] Laura Verde, Giuseppe De Pietro, and Giovanna Sannino. Voice Disorder Identification by using Machine Learning Techniques. *IEEE access*, 6:16246–16255, 2018.

[48] Shovan Bhowmik, Mahedi Hasan, and Muhammad Ataul Hakim. A dimensionality reduction based efficient multiple voice disease recognition scheme using mel-frequency cepstral coefficients and k-nearest neighbors algorithm. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*, pages 301–313. Springer, 2021.

[49] WJ Barry and M Putzer. Saarbrucken Voice Database. Institute of Phonetics, University of Saarland, 2007.

[50] Ziqi Fan, Jinyang Qian, Baoyin Sun, Di Wu, Yishen Xu, and Zhi Tao. Modeling Voice Pathology Detection using Imbalanced Learning. In *International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, pages 330–334. IEEE, 2020.

[51] Ziqi Fan, Yuanbo Wu, Changwei Zhou, Xiaojun Zhang, and Zhi Tao. Class-imbalanced Voice Pathology Detection and Classification using Fuzzy Cluster Oversampling Method. *Applied Sciences*, 11(8):3450, 2021.

[52] Özkan Arslan. A machine learning approach for voice pathology detection using mode decomposition-based acoustic cepstral features. *Mathematical Modelling and Numerical Simulation with Applications*, 4(4):469–494, 2024.

[53] S Syed, Munaf Rashid, Samreen Hussain, Anoshia Imtiaz, Hamnah Abid, and Hira Zahid. Inter Classifier Comparison to Detect Voice Pathologies. *Mathematical Biosciences and Engineering*, 18(3):2258–2273, 2021.

[54] Mounira Chaiani, Sid Ahmed Selouani, Malika Boudraa, and Mohammed Sidi Yakoub. Voice disorder classification using speech enhancement and deep learning models. *Biocybernetics and Biomedical Engineering*, 42(2):463–480, 2022.

[55] Fahad Taha Al-Dhief, Nurul Mu'azzah Abdul Latiff, Nik Noordini Nik Abd Malik, Naseer Sabri Salim, Marina Mat Baki, Musatafa Abbas Abbood Albadr, and Mazin Abed Mohammed. A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. *IEEE Access*, 8:64514–64533, 2020.

[56] Mujeeb Ur Rehman, Arslan Shafique, Sajjad Shaukat Jamal, Youcef Gheraibia, Aminu Bello Usman, et al. Voice Disorder Detection using Machine Learning Algorithms: An Application in Speech and Language Pathology. *Engineering Applications of Artificial Intelligence*, 133:108047, 2024.

[57] Pak Ho Leung, Kwok Tai Chui, Kenneth Lo, and Patricia Ordóñez de Pablos. A Support Vector Machine–based Voice Disorders Detection Using Human Voice Signal. In *Artificial Intelligence and Big Data Analytics for Smart Healthcare*, pages 197–208. Elsevier, 2021.

[58] Huiyi Wu, John Soraghan, Anja Lowit, and Gaetano Di Caterina. A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks. In *Interspeech 2018*, 2018.

[59] Vikas Mittal and RK Sharma. Deep Learning Approach for Voice Pathology Detection and Classification. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 16:1–30, 2021.

[60] Onur Karaman, Hakan Çakın, Adi Alhudhaif, and Kemal Polat. Robust Automated Parkinson Disease Detection Based on Voice Signals with Transfer Learning. *Expert Systems with Applications*, 178:115013, 2021.

[61] Xin Ning, Xinran Wang, Shaohui Xu, Weiwei Cai, Liping Zhang, Lina Yu, and Wenfa Li. A review of research on co-training. *Concurrency and computation: practice and experience*, 35(18):e6276, 2023.

[62] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[63] Wei Zhang, Qun Zhao, Yayu Liu, and Minhui Pang. Co-training approach for label-minimized audio classification. In *2010 International Conference on Measuring Technology and Mechatronics Automation*, volume 1, pages 860–863. IEEE, 2010.

[64] Md Mezbaur Rahman and Cornelia Caragea. Llm-guided co-training for text classification. *arXiv preprint arXiv:2509.16516*, 2025.

[65] MY Ge, CC Yu, L Zhou, and Y Ma. Deep learning image classification algorithm based on semi-supervised collaboration training. *Comput Simul*, 36(2):206–210, 2019.

[66] Jingdong Yang, Haoqiu Li, Han Wang, and Man Han. 3d medical image segmentation based on semi-supervised learning using deep co-training. *Applied Soft Computing*, 159:111641, 2024.

[67] Zhiqiang Shen, Peng Cao, Hua Yang, Xiaoli Liu, Jinzhu Yang, and Osmar R Zaiane. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. *arXiv preprint arXiv:2301.04465*, 2023.

[68] Yufan Yi, Yiping Xu, Bo Yang, and Yan Tian. A weighted co-training framework for emotion recognition based on eeg data generation using frequency-spatial diffusion transformer. *IEEE Transactions on Affective Computing*, 2024.

[69] Jelena Slivka and Aleksandar Kovačević. Co-training based algorithm for gender detection from emotional speech.

[70] Ya Hui Zhao, Hong Li Wang, and Rong Yi Cui. Abnormal voice detection algorithm based on semi-supervised co-training algorithm. *Advanced Materials Research*, 461:117–122, 2012.

[71] Carlos Guerrón Pantoja, Edgar Maya-Olalla, Hernán M Domínguez-Limaico, Marcelo Zambrano, Carlos Vásquez Ayala, and Marco Gordillo Pasquel. A Voice-Based Emotion Recognition System Using Deep Learning Techniques. In *International Conference on Innovation and Research*, pages 155–172. Springer, 2023.

[72] Ahmed Ismail, Samir Abdlerazek, and Ibrahim M El-Henawy. Development of Smart Healthcare System Based on Speech Recognition Using Support Vector Machine and Dynamic Time Warping. *Sustainability*, 12(6):2403, 2020.

[73] Ahmed Al-Nasheri, Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Khalid H Malki, Tamer A Mesallam, and Mohamed Farahat Ibrahim. Voice Pathology Detection and Classification Using Auto-Correlation and Entropy Features in Different Frequency Regions. *Ieee Access*, 6:6961–6974, 2017.

[74] Miklós Gábriel Tulics and Klára Vicsi. The Automatic Assessment of the Severity of Dysphonia. *International Journal of Speech Technology*, 22:341–350, 2019.

[75] RT Ritchings, Mark McGillion, and CJ Moore. Pathological Voice Quality Assessment using Artificial Neural Networks. *Medical engineering & physics*, 24(7-8):561–564, 2002.

Research candidate **Research supervisor**

Sharal Coelho
Research Scholar,
Dept. of Computer Science,
Mangalore University.